

12-001

Government
Publications



SURVEY METHODOLOGY



Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 2003

•

VOLUME 29

•

NUMBER 2



Statistics
Canada

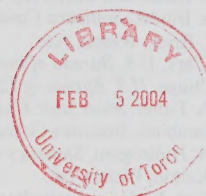
Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA



DECEMBER 2003 • VOLUME 29 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2004

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

January 2004

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members D.A. Binder

G.J.C. Hole

C. Patrick

R. Platek (Past Chairman)

E. Rancourt (Production Manager)

D. Roy

M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*

D.A. Binder, *Statistics Canada*

J.M. Brick, *Westat, Inc.*

C. Clark, *U.S. Bureau of the Census*

J. Eltinge, *U.S. Bureau of Labor Statistics*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistics Canada*

M.A. Hidirolou, *Statistics Canada*

G. Kalton, *Westat, Inc.*

P. Kott, *National Agricultural Statistics Service*

P. Lahiri, *JPSM, University of Maryland*

S. Linacre, *Australian Bureau of Statistics*

G. Nathan, *Hebrew University, Israel*

D. Norris, *Statistics Canada*

D. Pfeffermann, *Hebrew University*

J.N.K. Rao, *Carleton University*

T.J. Rao, *Indian Statistical Institute*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

R. Sitter, *Simon Fraser University*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

R. Valliant, *JPSM, University of Michigan*

J. Waksberg, *Westat, Inc.*

K.M. Wolter, *Iowa State University*

A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, H. Mantel and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts prepared following the guidelines given in the Journal in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. E-mail: singhmp@statcan.ca. Four nonreturnable printed copies of each manuscript can also be sent.

Subscription Rates

The price of *Survey Methodology* (Catalogue no. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$30 (\$15 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 29, Number 2, December 2003

CONTENTS

In This Issue	105
---------------------	-----

Discussion Paper


J.N.K. RAO, A.J. SCOTT and E. BENHIN Undoing Complex Survey Data Structures: Some Theory and Applications of Inverse Sampling	107
Comment: JOHN L. ELTINGE	119
SUSAN HINKINS	122
Response from the authors	126

Special Section on Census Coverage Error

HOWARD HOGAN The Accuracy And Coverage Evaluation: Theory and Design	129
PATRICK J. CANTWELL and MICHAEL IKEDA Handling Missing Data in the 2000 Accuracy and Coverage Evaluation Survey	139
H. ÖZTAŞ AYHAN and SÜHENDAN EKNI Coverage Error in Population Censuses: The Case of Turkey	155
D. COCCHI, E. FABRIZI and C. TRIVISANO A Hierarchical Model for the Analysis of Local Census Undercount in Italy	167

Regular Papers

C.J. SKINNER and R.G. CARTER Estimation of a Measure of Disclosure Risk for Survey Microdata Under Unequal Probability Sampling	177
J.P. REITER Inference for Partially Synthetic, Public Use Microdata Sets	181
K.R.W. BREWER and MARTIN E. DONADIO The High Entropy Variance of the Horvitz-Thompson Estimator	189
MOSUK CHOW and STEVEN K. THOMPSON Estimation with Link – Tracing Sampling Designs A Bayesian Approach	197
Acknowledgements	207



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743704>

In This Issue

This issue of *Survey Methodology* includes a special section on Census Coverage Error which presents four papers, including two papers on the coverage survey used in the United States, one from Turkey, and one from Italy. The special section is preceded by a discussed paper, and followed by four papers on various topics.

In the first paper of this issue, Rao, Scott and Benhin study the repeated inverse sampling method proposed by Hinkins, Oh and Scheuren. In this approach, random subsamples are drawn from a complex sample in such a way that each subsample is unconditionally a simple random sample from the population. Rao, Scott and Benhin present some theoretical results for the expectation and variance of the repeated inverse sampling estimator. They then explore some conditions under which the repeated inverse sampling estimator converges to the original full sample estimator. They finally propose an approach based on estimating equations that avoids some of the potential bias of the repeated inverse sampling estimator for nonlinear parameters. The paper is followed by two fascinating discussions by Eltinge and Hinkins, and a rejoinder by the authors.

Hogan, in the first paper of the special section on Census Coverage Errors, presents a concise overview of the survey used to provide estimates for net undercoverage in the 2000 Census. He presents the Accuracy and Coverage Evaluation (ACE) study in the context of general post enumeration surveys and dual system estimators. He also presents the assumptions needed for these types of surveys to produce unbiased estimates and a detailed discussion where these assumptions failed in the 2000 ACE. The results are very interesting.

The next paper is also concerned with the 2000 ACE. Cantwell and Ikeda examine the crucial assumptions made when some data is missing. One of the points the authors note is that when a rare characteristic – persons missed by the Census in this case – is being estimated the methods used to adjust for missing data are very important. The authors point out the changes made from the methods used in previous post enumeration surveys for the 2000 ACE.

Ayhan and Ekni present the coverage procedures used in a different census context. While the basic post enumeration survey design is used in Turkey, there are some interesting differences between their experiences and those of the United States. Since Turkey uses a de facto approach to Census residence as opposed to the de jure approach used in the United States, there are some operational differences in the post enumeration surveys. These are clearly pointed out by the authors.

The final paper in the special section on Census Coverage Errors, by Cocchi, Fabrizi and Trivisano, describes the 1991 Italian Population Census and the the Post Enumeration Survey (PES) used to measure undercount. Since the census is administered by municipalities, data on the statistical quality of municipalities are used as auxiliary information for PES modelling and estimation. Poisson regression trees and hierarchical Poisson models are used to analyze the data. Results are summarized and discussed, and some recommendations are given.

Skinner and Carter extend estimation for Skinner and Elliot's measure of disclosure risk for survey microdata from the equal probability sampling case to the unequal probability sampling case under an assumption of Poisson sampling. Effects of possible departures from Poisson sampling are also considered.

The problem of inference for partially synthetic microdata sets is considered by Reiter. Statistical agencies may release microdata sets with completely synthetic data in order to protect respondent confidentiality. Methods for inference when the complete dataset is synthetic have been developed but most agencies release only partially synthetic datasets, that is, datasets for which only sensitive variables are imputed. There has been little reported in the literature under this situation. Reiter's proposed method is shown to be valid under a Bayesian framework and under a design-based framework and is illustrated by simulation studies.

In Brewer and Donadio, a variance estimator for the Horvitz-Thompson estimator that does not require the calculation of the second-order inclusions probabilities is obtained under high entropy situations. High entropy situations occur when there is the absence of any detectable pattern or ordering in the selected sample units. Under high entropy situations, an approximate variance formula is derived and verified through a model-assisted approach. A sample estimator for this approximate design-variance of the Horvitz-Thompson estimator is then developed. Finally, the proposed estimator is empirically compared with several other estimators using several populations.

Finally, Chow and Thompson present a Bayesian approach to designs where social links are exploited to obtain a sample of hidden or hard-to-access human populations. The authors provide an accessible introduction to the Bayesian approach in which the social links from one person to another are used to create the prior distribution. It is easy to adjust these priors when information is vague. The result is that from the resulting posterior distribution a large number of questions can be answered.

M.P. Singh

Undoing Complex Survey Data Structures: Some Theory and Applications of Inverse Sampling

J.N.K. RAO, A.J. SCOTT and E. BENHIN¹

ABSTRACT

Application of classical statistical methods to data from complex sample surveys without making allowance for the survey design features can lead to erroneous inferences. Methods have been developed that account for the survey design, but these methods require additional information such as survey weights, design effects or cluster identification for microdata. Inverse sampling (Hinkins, Oh and Scheuren 1997) provides an alternative approach by undoing the complex survey data structures so that standard methods can be applied. Repeated subsamples with unconditional simple random sampling structure are drawn and each subsample analysed by standard methods and then combined to increase the efficiency. This method has the potential to preserve confidentiality of microdata, although computer-intensive. We present some theory of inverse sampling and explore its limitations. A combined estimating equations approach is proposed for handling complex parameters such as ratios and "census" linear regression and logistic regression parameters. The method is applied to a cluster correlated data set reported in Battese, Harter and Fuller (1988).

KEY WORDS: Combined estimating equations; Confidentiality; Repeated subsampling.

1. INTRODUCTION

There is a fairly clear distinction between the focus of traditional sample survey methodology and that of the rest of applied statistics. Survey samplers have concentrated on developing efficient (but complicated) ways of drawing samples to estimate rather simple quantities (population means, proportions, totals, *etc.*). Most other applied statisticians, by contrast, have concentrated on developing sophisticated methods for fitting very complicated models, but assuming a rather simple sampling structure (often that the observations are independent).

In reality, data from complicated surveys are often used to fit complicated models. For example, people may want to use data from a Labour Force Survey to characterize the association between education and unemployment levels. They might want to use data from health surveys to study the association between housing conditions or poverty and morbidity, and so on. Extending the range of application of standard methods so that they can be applied to data from complicated sample surveys, involving multi-stage sampling and variable selection probabilities, is difficult and cumbersome; see *e.g.*, Skinner, Holt and Smith (1989).

How do practitioners deal with the complexity of survey data structures? Adapting a quote from Hinkins, Oh and Scheuren (1997) (abbreviated HOS hereafter): "If your only tool is a hammer, every problem looks like a nail!"; the hammer available to most people is one of the big statistical packages (SAS, Splus, SPSS, *etc.*). Most people still just push their data through a standard program and ignore the survey design features. This is in spite of the fact that a

great deal of effort over the last two decades has been spent on developing methods to analyze survey data that take account of design features, and specialized programs such as SUDAAN or WesVar are now available to implement some of these methods.

An alternative to developing complex new tools (which may rarely be used in practice anyway!) is to work backwards: instead of tailoring the methods to fit the data, tailor the data to fit the methods. One approach along these lines was developed in Rao and Scott (1992; 1999). Another approach has been suggested in HOS. Their basic idea is to avoid the pain caused by a complicated sample by choosing a subsample (inverse-sample) that has a simple random sample structure unconditionally (or at least has a structure that is considerably simpler to handle than the original sample). Obviously this involves some loss in efficiency, especially if the subsample is very much smaller than the original sample, as often turns out to be necessary. However, we can increase the efficiency by repeating the process independently many times and averaging the results.

Is it possible to produce subsamples with the desired properties? The answer is often "yes", although the resulting subsample size, m , might have to be small (in fact, no more than $m = 2$ for some standard stratified multistage designs). HOS give algorithms for producing simple random inverse-samples for a number of standard designs. We summarize the inverse sampling schemes in section 2 for ready reference. These schemes include both exact and approximate methods in terms of matching simple random sampling. In this paper we look at some of the properties of

¹ J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Canada, K1S 5B6. E-mail: jrao@math.carleton.ca; A.J. Scott, Department of Statistics, University of Auckland, Auckland, New Zealand. E-mail: scott@stat.auckland.ac.nz; E. Benhin, Household Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6. E-mail: emmanuel.benhin@statcan.ca.

the repeated inverse sampling procedures given in section 2. In particular, we develop some basic theory of inverse sampling in section 3, and illustrate some of the strengths and weaknesses of the procedure. In section 4, we study the special case of a population total. We propose a combined estimating equations (CEE) approach in section 5 for handling complex parameters such as ratios and "census" regression parameters. Finally, some concluding remarks are given in section 6. Proofs of theorems are given in the appendix.

2. INVERSE SAMPLING ALGORITHMS

In this section we summarize the inverse sampling schemes, proposed by Hinkins *et al.* (1997), for ready reference. These schemes include both exact and approximate methods in terms of matching simple random sampling (SRS) unconditionally.

Suppose we have a sample s_0 of observations drawn from the finite population of size N according to a specified complex design. We wish to draw a subsample s^* of size m from s_0 such that the unconditional probability of s^* , $p(s^*)$, matches simple random sampling with $p(s^*) = 1 / \binom{N}{m}$, either exactly or approximately. We have

$$p(s^*) = \sum_{s_0 \supset s^*} p_0(s_0) p(s^* | s_0), \quad (2.1)$$

where $p_0(s_0)$ is the probability of selecting s_0 and $p(s^* | s_0)$ is the conditional probability of choosing s^* . If $p(s^* | s_0)$ does not depend on s_0 , then it follows from (2.1) that

$$p(s^* | s_0) = p_2(s^*) = \frac{p(s^*)}{\sum_{s_0 \supset s^*} p_0(s_0)}. \quad (2.2)$$

Denote the first-order and second-order inclusion probabilities corresponding to s^* and s_0 as (π_i^*, π_{il}^*) and (π_i, π_{il}) respectively, where $\pi_i^* = m/N$ and $\pi_{il}^* = m(m-1)/(N(N-1))$, $i \neq l$. Similarly, denote the conditional inclusion probabilities as $(\tilde{\pi}_i(s_0), \tilde{\pi}_{il}(s_0))$. If the conditional inclusion probabilities do not depend on s_0 , then we write them as $(\tilde{\pi}_i, \tilde{\pi}_{il})$. It is readily seen that

$$\pi_i^* = \sum_{s_0 \ni i} p_0(s_0) \tilde{\pi}_i(s_0); \quad \pi_{il}^* = \sum_{s_0 \ni i, l} p_0(s_0) \tilde{\pi}_{il}(s_0). \quad (2.3)$$

If $\tilde{\pi}_i(s_0) = \tilde{\pi}_i$ and $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il}$, then it follows from (2.3) that

$$\pi_i^* = \pi_i \tilde{\pi}_i, \quad \pi_{il}^* = \pi_{il} \tilde{\pi}_{il}. \quad (2.4)$$

In section 4 we use (2.4) to study the properties of inverse sampling for estimating a population total. Note that (π_i^*, π_{il}^*) may correspond to some other simpler sampling design if it is not feasible to match simple random sampling (SRS), e.g., stratified simple random sampling.

2.1 Stratified Simple Random Sampling

Suppose that the original sample s_0 is a stratified simple random sample, i.e.,

$$p_0(s_0) = \prod_{h=1}^L \binom{N_h}{n_h}^{-1}, \quad (2.5)$$

where $N_h(n_h)$ denotes the number of population (sample) units in stratum h ($= 1, \dots, L$). We wish to draw a subsample s^* of size m such that $p(s^*) = 1 / \binom{N}{m}$, where $N = \sum_{h=1}^L N_h$. Clearly, m cannot be larger than $\min(n_h)$. Let $\mathbf{m} = (m_1, \dots, m_L)^T$ denote the (random) number of units in each stratum that belong to s^* , $0 \leq m_h \leq m$, $\sum_{h=1}^L m_h = m$. Noting that the number of terms in $\sum_{s_0 \supset s^*}$ equals $\prod_{h=1}^L \binom{N_h - m_h}{n_h - m_h}$, it follows from (2.2) that

$$p(s^* | s_0) = \frac{\prod_{h=1}^L \binom{N_h}{m_h}}{\binom{N}{m}} \frac{1}{\prod_{h=1}^L \binom{n_h}{m_h}}. \quad (2.6)$$

The subsampling scheme readily follows from (2.6): (i) Generate \mathbf{m} from the hypergeometric distribution $f(\mathbf{m}) = \prod_{h=1}^L \binom{N_h}{m_h} / \binom{N}{m}$; (ii) Draw a simple random sample of size m_h , without replacement, from the n_h sample units in stratum h , independently across strata h ($= 1, \dots, L$). HOS specify $p(s^* | s_0)$ first and then verify that it gives $p(s^*) = \binom{N}{m}^{-1}$. Our approach provides the subsampling scheme from the specification of $p_0(s_0)$ and $p(s^*)$.

2.2 One-stage Cluster Sampling

HOS studied the case of one-stage cluster sampling in detail. Three sampling designs for s_0 were investigated: (1) Equal cluster sizes, M , and simple random sampling of clusters; (2) Unequal cluster sizes, M_i , and simple random sampling of clusters; (3) Unequal cluster sizes, M_i , and clusters sampled with probability proportional to size M_i and with replacement.

Case 1. Exact matching with SRS is difficult to implement in the case of equal cluster sizes, M , and simple random sampling of clusters. Suppose s_0 contains k clusters drawn from K clusters in the population ($N = KM$). A simple approximate method of subsampling selects one element at random from each sample cluster so that the size of s^* is k . Hoffman, Sen and Weinberg (2001) used a similar method for biostatistical applications. HOS used systematic sampling to select one case from each sample cluster.

Case 2. Hoffman *et al.* (2001) selected one unit at random from each cluster in the case of unequal cluster sizes, under a model-based framework for clustered data. For sampling applications, this method does not work in the sense that it is not possible to obtain SRS of fixed sizes by subsampling, even approximately. HOS proposed an alternative method that artificially enlarges the population to equal cluster size case and then applies subsampling used in Case 1. We first force all clusters to have the same size by adding an

appropriate number of pseudo-unit to bring them up to the size of the largest sample cluster. Then we take one unit at random from each sample cluster, and discard any pseudo-units to obtain the final sample. This approximate method makes $p(s^*|s_0)$ depend on s_0 because the conditional probability depends on $M(s_0)$, the size of the largest sample cluster.

Case 3. For the case of unequal cluster sizes M_i and probability proportional to size (PPS) sampling with replacement, HOS proposed a simple method of subsampling which gives $p(s^*) = (1/N)^k$, where s^* now denotes an ordered simple random sample drawn with replacement from the $N = \sum_{i=1}^K M_i$ units in the population, i.e., $s^* = (i_1, \dots, i_j, \dots, i_k)$, where i_j denotes the unit drawn in the j -th draw ($j = 1, \dots, k$). Viewing the sample clusters as ordered, we select one unit at random from each sample cluster. Note that the same cluster might appear more than once in the ordered sample. Denote the size of the cluster drawn in the i -th PPS draw by M'_i , then

$$p(s^*) = \left[\prod_{i=1}^k \frac{M'_i}{N} \right] \left[\prod_{i=1}^k \frac{1}{M'_i} \right] = \left(\frac{1}{N} \right)^k, \quad (2.7)$$

where $\prod_{i=1}^k (M'_i/N)$ is the probability of drawing the ordered cluster sample. Note that s_0 is the ordered PPS sample and we have only one term in the summation in (2.1).

If the clusters are drawn with inclusion probabilities $\pi_i = kM_i/N$ and without replacement, then it is not possible to match SRS. However, we can treat the clusters as if they were drawn with replacement, as done in practice, and then apply the scheme for Case 3. This will lead to overestimation of variance if the variance of the estimator is smaller than the variance of the estimator under PPS sampling with replacement (see e.g., Wolter 1985, page 45). However, the overestimation is not serious if the sampling fraction k/K is small (see Section 4.3).

2.3 Two-stage Cluster Sampling

HOS also studied two-stage sampling for the following cases: (1) Equal cluster sizes, M , and k clusters sampled with equal probability in the first stage; simple random subsample of equal size, m , drawn independently within each sampled cluster (PSU). (2) Unequal cluster sizes, M_i , and k clusters sampled with PPS and with replacement; simple random subsamples of unequal sizes, m_i , drawn independently within each cluster in the with replacement sample.

Case 1. As in the case of one-stage cluster sampling, exact method of inverse sampling is difficult to implement. A simple approximate method of inverse sampling selects one unit at random from each of the k subsamples.

Case 2. As in Case 3 of uni-stage cluster sampling, we simply select one unit at random from each of the ordered subsamples. HOS suggested a different method: Take a simple random sample with replacement of k clusters first and then with each selected cluster take one unit at random from the corresponding subsample. It appears that the first stage inverse sampling of clusters is not necessary. To see this, we note that

$$p_0(s_0) = \prod_{i=1}^k \left[\left(\frac{M'_i}{N} \right) \frac{1}{\binom{M'_i}{m'_i}} \right],$$

where m'_i is the subsample size associated with the cluster selected in the i -th draw ($i = 1, \dots, k$). We wish to draw a subsample s^* of size k such that $p(s^*) = (1/N)^k$, where $N = \sum_{i=1}^K M_i$. Also the number of terms in $\sum_{s_0 \supset s^*}$ equals $\prod_{i=1}^k \binom{M'_i - 1}{m'_i - 1}$ and

$$\sum_{s_0 \supset s^*} p_0(s_0) = \prod_{i=1}^k \left[\left(\frac{M'_i}{N} \right) \frac{\binom{M'_i - 1}{m'_i - 1}}{\binom{M'_i}{m'_i}} \right] = \prod_{i=1}^k \frac{m'_i}{N}.$$

It follows from (2.2) that $p(s^*|s_0) = \prod_{i=1}^k (1/m'_i)$ and hence the subsampling scheme readily follows.

2.4 Stratified Two-stage Sampling

Suppose we have a two-stage sample from each stratum, where the clusters are sampled with PPS with replacement and subsampling is done independently within each sampled cluster. Using the inverse sampling procedure of Case 2, section 2.3, we get simple random samples from each stratum. We can then apply the method of section 2.1, treating the inverse-samples as if drawn without replacement to get an inverse-sample of size $k_0 = \min_h (k_h)$, where k_h is the number of sampled clusters in stratum h . In the important case of $k_h = 2$ psu's sampled from each stratum, the inverse-sample size, k_0 , is only two.

3. BASIC PROPERTIES

The results in this section are quite general and apply equally to sample surveys and the type of clustered situation considered by Hoffman *et al.* (2001). Suppose that we are interested in estimating some population parameter, θ , and we have a sample, s_0 , of observations drawn from the population according to some complex design. We assume that we have a subsampling algorithm that can produce samples from some simpler design. This design will often be simple random sampling, but we can extend the range of

applications considerably by allowing for the possibility of more general (sub-)designs; for example, stratified SRS when the original sample is a stratified two-stage sample. Our only requirement for the simpler design is that we can produce an estimator of the quantity of interest, θ , together with an estimator of its variance. Let $\hat{\theta}_j^*$ and \hat{V}_j^* denote the estimator and variance estimator produced from the j -th subsample when we generate a sequence of g conditionally independent subsamples s_j^* ($j = 1, \dots, g$). Note that the $\hat{\theta}_j^*$'s are not unconditionally independent when averaged over the distribution of the initial sample, s_0 . A "separate" inverse-sampling estimator of θ based on the g subsamples is given by

$$\hat{\theta}_g = \frac{1}{g} \sum_{j=1}^g \hat{\theta}_j^*. \quad (3.1)$$

We denote the estimator based on s_0 as $\hat{\theta}$. Theorem 1 below gives basic results on $\hat{\theta}_g$ and its variance.

Theorem 1

1. Conditional on the original sample, s_0 , $\hat{\theta}_g$ converges almost surely to $E(\hat{\theta}_1^* | s_0) = \hat{\theta}_\infty$, say, as $g \rightarrow \infty$.
2. $E(\hat{\theta}_g) = E(\hat{\theta}_1^*)$.
3. $\text{Var}(\hat{\theta}_g) = \text{Var}(\hat{\theta}_\infty) + \frac{1}{g} E[\text{Var}(\hat{\theta}_1^* | s_0)]$.
4. If $r_g = \frac{\text{Var}(\hat{\theta}_g)}{\text{Var}(\hat{\theta}_\infty)}$, then $r_g = 1 + \frac{r_1 - 1}{g}$.

Result 4 of Theorem 1 demonstrates that increasing the number of subsamples, g , does indeed increase the efficiency of $\hat{\theta}_g$. More precisely, the variance ratio r_g has the form $a + b/g$. If the subsample estimator, $\hat{\theta}_1^*$, is unbiased for θ , then so is the inverse-sampling estimator, $\hat{\theta}_g$. However, if $\hat{\theta}_1^*$ has bias of order m^{-1} , where m denotes the subsample size, then $\hat{\theta}_g$ has exactly the same bias. Since m will usually be very much smaller than the original sample size, this bias can be appreciable. This is a serious limitation of $\hat{\theta}_g$ in the nonlinear cases, such as ratios and regression coefficients. In section 5, we propose an alternative estimator of θ based on the estimating equations (EE) approach. This estimator is asymptotically unbiased for any m as the size of s_0 increases, unlike $\hat{\theta}_g$.

Result 4 of Theorem 1 can be used to determine the number of subsamples, g needed to obtain reasonable efficiency. For example, HOS give an example in which $r_1 = 29.3$. The original sample was a very efficient stratified random sample with $n = 15,618$ observations taken from the Statistics of Income corporate survey, while the subsample was a simple random sample of $m = 2,224$ observations. A single subsample is relatively inefficient. However, in this case, repeated inverse sampling recovers all the information in the original sample in the limit. Applying Result 4 of Theorem 1 leads immediately to the following table:

g	1	10	100	1000
r_g	29.3	3.83	1.28	1.03

(HOS produced these same results by simulation but this is unnecessary in view of Result 4 of Theorem 1.) We see that $g = 100$ subsamples would be adequate for many purposes and that we obtain almost full efficiency with $g = 1,000$.

The fact that $\hat{\theta}_1^*, \dots, \hat{\theta}_g^*$ are not unconditionally independent means that estimating $\text{Var}(\hat{\theta}_g)$ is not completely straightforward. However, a relatively simple variance estimator may be obtained using Theorem 2 below.

Theorem 2

The variance of $\hat{\theta}_g$ may be expressed as

$$\text{Var}(\hat{\theta}_g) = \text{Var}(\hat{\theta}_1^*) - \frac{g-1}{g} E[\text{Var}(\hat{\theta}_1^* | s_0)]. \quad (3.2)$$

We can estimate the first term of (3.2) by \hat{V}_j^* for $j = 1, \dots, g$, and hence by their average $g^{-1} \sum_{j=1}^g \hat{V}_j^*$. In addition, the quantity

$$s_{\theta g}^2 = \frac{1}{g-1} \sum_{j=1}^g (\hat{\theta}_j^* - \hat{\theta}_g)^2$$

gives an unbiased estimator of $E[\text{Var}(\hat{\theta}_1^* | s_0)]$ because $\hat{\theta}_1^*, \dots, \hat{\theta}_g^*$ are conditionally independent given the initial sample, s_0 . This leads to an estimator of $\text{Var}(\hat{\theta}_g)$ of the form

$$\hat{V}_g = \frac{1}{g} \sum_{j=1}^g \hat{V}_j^* - \frac{1}{g} \sum_{j=1}^g (\hat{\theta}_j^* - \hat{\theta}_g)^2. \quad (3.3)$$

The properties of the variance estimator \hat{V}_g depend on the properties of the subsample estimator \hat{V}_j^* . For example, if \hat{V}_j^* is unbiased, then \hat{V}_g is also unbiased.

For the special case of a population total $\theta = Y$ and simple random subsampling, i.e., $p(s^*) = 1/\binom{N}{m}$, we have $\hat{\theta}_j^* = \hat{Y}_j^* = N\bar{y}_j^*$ and \hat{V}_j^* is unbiased for Y with unbiased variance estimator $\hat{V}_j^* = N^2(m^{-1} - N^{-1})s_{jy}^{*2}$, where \bar{y}_j^* is the mean and s_{jy}^{*2} is the variance of the j -th subsample. The variance estimator \hat{V}_g of $\hat{\theta}_g = \hat{Y}_g = g^{-1} \sum_{j=1}^g (N\bar{y}_j^*)$, given by (3.3), is unbiased, and it reduces to

$$\hat{V}_g = \frac{1}{g} \sum_{j=1}^g \hat{V}_j^* - \frac{N^2}{g} \sum_{j=1}^g (\bar{y}_j^* - \bar{y}_g)^2, \quad (3.4)$$

where $\bar{y}_g = g^{-1} \sum_{j=1}^g \bar{y}_j^*$. HOS derived a variance estimator by first expressing $\text{Var}(\hat{Y}_g)$ as

$$\begin{aligned} \text{Var}(\hat{Y}_g) &= N^2 \frac{m-1}{m} S_y^2 + \frac{1}{g} \sum_{j=1}^g \text{Var}(\hat{Y}_j^*) \\ &\quad - N^2 \frac{mg-1}{mg} E[s_{cy}^{*2}], \end{aligned} \quad (3.5)$$

where S_y^2 is the population variance and s_{cy}^{*2} is the sample variance using all gm subsample units. In the second step, they remarked that we can generate an approximately unbiased estimator of $\text{Var}(\hat{Y}_g)$ from (3.5) by replacing S_y^2 and $\text{Var}(\hat{Y}_j^*)$ with unbiased estimators and replacing $E(s_{cy}^{*2})$ by s_{cy}^{*2} . We now follow this recipe and obtain an explicit form for the HOS variance estimator, denoted $\hat{V}_{g(\text{HOS})}$. Noting that each s_{jy}^{*2} is unbiased for S_y^2 , a pooled unbiased estimator of S_y^2 is obtained as $g^{-1} \sum_{j=1}^g s_{jy}^{*2}$. Further, s_{cy}^{*2} may be decomposed as $(mg-1)s_{cy}^{*2} = (m-1) \sum_{j=1}^g s_{jy}^{*2} + m \sum_{j=1}^g (\bar{y}_j^* - \bar{y}_g)^2$. Hence,

$$\begin{aligned} \hat{V}_{g(\text{HOS})} &= N^2 \left(\frac{m-1}{m} \right) \frac{1}{g} \sum_{j=1}^g s_{jy}^{*2} + \frac{1}{g} \sum_{j=1}^g \hat{V}_j^* \\ &\quad - N^2 \left\{ \left(\frac{m-1}{m} \right) \frac{1}{g} \sum_{j=1}^g s_{jy}^{*2} + \frac{1}{g} \sum_{j=1}^g (\bar{y}_j^* - \bar{y}_g)^2 \right\} \\ &= \hat{V}_{g^*} \end{aligned} \quad (3.6)$$

It follows from (3.6) that the variance estimator of HOS is in fact identical to our variance estimator (3.4) and also exactly unbiased.

4. ESTIMATION OF A TOTAL

4.1 Exact Matching

As shown in section 3, repeated subsampling increases the efficiency of an estimator, but this does not necessarily mean that the inverse-sampling estimator, $\hat{\theta}_g$, converges to the original full sample estimator, $\hat{\theta}$, as $g \rightarrow \infty$, even when we start with an unbiased estimator for the subsample. In this section, we study the special case of a total $\theta = Y$ and consider the Horvitz-Thompson (H-T) unbiased estimator, $\hat{Y} = \sum_{i \in s_0} y_i / \pi_i$, based on the original full-sample. Theorem 3 below establishes conditions under which the corresponding inverse-sampling estimator

$$\hat{Y}_g = \frac{1}{g} \sum_{j=1}^g \hat{Y}_j^* \quad (4.1)$$

converges to the H-T estimator, \hat{Y} , for the original design as $g \rightarrow \infty$, where

$$\hat{Y}_j^* = \sum_{i \in s_j^*} \frac{y_i}{\pi_i^*}$$

and π_i^* is the unconditional inclusion probability for the i -th unit. If the subsample s_j^* is a simple random sample unconditionally, then $\pi_i^* = m/N$, where m is the subsample size.

Theorem 3

Let $\tilde{\pi}_i(s_0)$ be the conditional probability that the i -th unit is selected in the subsample for a given initial sample, s_0 .

Suppose that $\hat{\theta}_j^* = \hat{Y}_j^*$ is the H-T estimator of a total $\theta = Y$ for the j -th subsample. Then the limiting inverse-sampling estimator, $\hat{\theta}_\infty^* = \hat{Y}_\infty^*$, will be the H-T estimator, \hat{Y} , for the original design if and only if the conditional inclusion probabilities $\tilde{\pi}_i(s_0)$ are constant for all s_0 containing the i -th unit, i.e., $\tilde{\pi}_i(s_0) = \pi_i$ for all $s_0 \supset i$.

The condition $\tilde{\pi}_i(s_0) = \pi_i$ is a fairly natural one for most sampling designs for which the H-T estimator is used. If the subsamples are all simple random samples of fixed size m , then the estimator for a subsample is simply $N\bar{y}_j^*$, which is the natural estimator under simple random sampling.

Theorem 4 below establishes conditions under which the inverse-sampling variance estimator, $\hat{V}_{g,\text{HT}}$, of \hat{Y}_g converges to \hat{V}_{HT} , the H-T variance estimator of the full-sample estimator \hat{Y} , as $g \rightarrow \infty$. We have

$$\hat{V}_{\text{HT}} = \sum_{i,l \in s_0} \frac{\pi_{il} - \pi_i \pi_l}{\pi_i \pi_l \pi_{il}} y_i y_l \quad (4.2)$$

(see Cochran 1977, page 261) and

$$\hat{V}_{g,\text{HT}} = \frac{1}{g} \sum_{j=1}^g \hat{V}_{j,\text{HT}}^* - \frac{1}{g} \sum_{j=1}^g (\hat{Y}_j^* - \hat{Y}_g)^2$$

with

$$\hat{V}_{j,\text{HT}}^* = \sum_{i,l \in s_j^*} \frac{\pi_{il}^* - \pi_i^* \pi_l^*}{\pi_i^* \pi_l^* \pi_{il}^*} y_i y_l, \quad (4.3)$$

where π_{il}^* is the unconditional joint inclusion probability for the i -th and l -th units. If the subsample s_j^* is a simple random subsample unconditionally, then $\pi_{il}^* = m(m-1)/[N(N-1)]$, $i \neq l$. Note that $\hat{V}_{j,\text{HT}}^*$ is the H-T variance estimator of \hat{Y}_j^* , and $\pi_{ii}^* = \pi_i^*$, $\pi_{ii} = \pi_i$.

Theorem 4

If $\hat{V}_{j,\text{HT}}^*$ is the Horvitz-Thompson (H-T) variance estimator of \hat{Y}_j^* for the j -th subsample, then conditional on s_0 , $\hat{V}_{g,\text{HT}}$, converges to the Horvitz-Thompson (H-T) variance estimator of \hat{Y} for the original design, as $g \rightarrow \infty$, if the conditional joint inclusion probabilities are constant for all s_0 containing a given pair (i, l) of units, i.e., $\tilde{\pi}_{il}(s_0) = \pi_{il}$ for all $s_0 \supset \{i, l\}$.

In Theorem 4 we considered the H-T variance estimator. But the Sen-Yates-Grundy (S-Y-G) variance estimator, \hat{V}_{SYG} , is often preferred over the H-T variance estimator, \hat{V}_{HT} , because it is more stable and several designs for which it is always nonnegative are known, while \hat{V}_{HT} frequently takes negative values (Cochran 1977, page 261). The S-Y-G variance estimator of \hat{Y} exists for fixed sample size designs and it is given by

$$\hat{V}_{\text{SYG}} = \sum_{i < l \in s_0} \frac{\pi_i \pi_l - \pi_{il}}{\pi_{il}} \left(\frac{y_i}{\pi_i} - \frac{y_l}{\pi_l} \right)^2, \quad (4.4)$$

for the full-sample design. Similarly, the S-Y-G variance estimator of \hat{Y}_j^* is

$$\hat{V}_{j, \text{SYG}}^* = \sum_{i \in s_j} \sum_{l \in s_j} \frac{\pi_i^* \pi_l^* - \pi_{il}^*}{\pi_i^* \pi_l^*} \left(\frac{y_i}{\pi_i^*} - \frac{y_l}{\pi_l^*} \right)^2. \quad (4.5)$$

The inverse-sampling variance estimator is given by

$$\hat{V}_{g, \text{SYG}} = \frac{1}{g} \sum_{j=1}^g \hat{V}_{j, \text{SYG}}^* - \frac{1}{g} \sum_{j=1}^g (\hat{Y}_j^* - \hat{Y}_g)^2. \quad (4.6)$$

Theorem 5 below shows that $\hat{V}_{g, \text{SYG}}$ does not converge to \hat{V}_{SYG} as $g \rightarrow \infty$, i.e., $\hat{V}_{\infty, \text{SYG}} \neq \hat{V}_{\text{SYG}}$. If the subsample is a simple random sample unconditionally, i.e., $\pi_i^* = m/N$ and $\pi_{il}^* = m(m-1)/[N(N-1)]$; $i \neq l$, then $\hat{V}_{j, \text{HT}}^* = \hat{V}_{j, \text{SYG}}^*$ and $\hat{V}_{\infty, \text{SYG}} = \hat{V}_{\infty, \text{HT}} = \hat{V}_{\text{HT}}$, the H-T variance estimator of \hat{Y} .

Theorem 5

The inverse-sampling variance estimator (4.6) does not converge to the S-Y-G variance estimator (4.4) as $g \rightarrow \infty$.

4.2 Exact Matching: PPS Estimates

(i) Unistage cluster sampling

For the case of PPS sampling with replacement of clusters with unequal sizes M_i , we have exact matching with SRS with replacement. The estimates of Y is given by $\hat{Y}_{\text{pps}} = (N/k) \sum_{i=1}^k \bar{Y}_i'$, where N is the total number of population elements and \bar{Y}_i' is the mean of the cluster selected on the i -th draw. The estimator \hat{Y}_{pps} is not equal to the H-T estimator of Y . The variance estimator of \hat{Y}_{pps} is given by

$$\hat{V}_{\text{pps}} = \frac{N^2}{k} \frac{1}{k-1} \sum_{i=1}^k \left(\bar{Y}_i' - \frac{1}{k} \sum_{i=1}^k \bar{Y}_i' \right)^2.$$

The inverse-sampling estimator corresponding to \hat{Y}_{pps} is given by $\hat{Y}_g = g^{-1} \sum_{j=1}^g \hat{Y}_j^*$, where \hat{Y}_j^* denotes the estimator of Y from the j -th inverse sample. It is easy to verify that $\hat{Y}_{\infty} = \hat{Y}_{\text{pps}}$, and $\hat{V}_{\infty} = \hat{V}_{\text{pps}}$, noting that $\hat{Y}_j^* = (N/k) \sum_{i \in s_j} y_i'$ where y_i' denotes the value of the element of an inverse-sample selected from the cluster in the i -th draw. Thus, inverse sampling preserves both the estimator and the variance estimator.

(ii) Two-stage cluster sampling

Turning to the case of unequal cluster sizes, M_i , we select the clusters with PPS and with replacement, and then draw simple random subsampling of equal size, m , independently within each cluster in the with-replacement sample. The estimator of Y is $\hat{Y}_{\text{pps}} = (N/k) \sum_{i=1}^k \bar{y}_i'$ where \bar{y}_i' is the sample mean of the cluster selected in the i -th draw. The variance estimator of \hat{Y}_{pps} is given by

$$\hat{V}_{\text{pps}} = \frac{N^2}{k} \frac{1}{k-1} \sum_{i=1}^k \left(\bar{y}_i' - \frac{1}{k} \sum_{i=1}^k \bar{y}_i' \right)^2.$$

The inverse-sampling estimator is given by $\hat{Y}_g = g^{-1} \sum_{j=1}^g \hat{Y}_j^*$, where $\hat{Y}_j^* = (N/k) \sum_{i \in s_j} y_i'$, and y_i' is

defined as above. It is easy to verify that $\hat{Y}_{\infty} = \hat{Y}_{\text{pps}}$ and $\hat{V}_{\infty} = \hat{V}_{\text{pps}}$. Thus, inverse sampling preserves both the estimator and the variance estimator.

4.3 Approximate Matching

In section 2 we noted that exact matching with SRS is difficult to implement when the original sampling design involves clusters. We proposed several approximate matching methods to overcome this difficulty. In this subsection we study the properties of the approximate matching methods.

4.3.1 Unistage Cluster Sampling

In section 2.2, Case 1, we considered the case of equal cluster sizes, M , and simple random sampling of clusters. The estimator of a total Y is given by $\hat{Y} = (K/k) \sum_{i=1}^k Y_i$, where Y_i is the i -th sample cluster total and K is the number of population clusters. The variance estimator of \hat{Y} is

$$\hat{V} = \frac{K^2}{k} \left(1 - \frac{k}{K} \right) \frac{1}{k-1} \sum_{i=1}^k \left[Y_i - \frac{1}{k} \sum_{i=1}^k Y_i \right]^2.$$

For inverse sampling, we proposed approximate matching by selecting one unit at random from each sample cluster, $i(i=1, \dots, k)$. The inverse-sampling estimator is given by $\hat{Y}_g = g^{-1} \sum_{j=1}^g \hat{Y}_j^*$ with $\hat{Y}_j^* = N \bar{y}_j^*$ denoting the estimator of total Y from the j -th inverse-sample. The inverse-sampling variance estimator, \hat{V}_g , is given by (3.4).

It is easy to verify that $\hat{Y}_{\infty} = \hat{Y}$ so that approximate matching preserves the original estimator \hat{Y} in the limit. On the other hand, it can be shown that

$$\hat{V}/\hat{V}_{\infty} = 1 - k/K. \quad (4.7)$$

It now follows from (4.7) that \hat{V}_{∞} leads to overestimation of the variance if the sampling fraction k/K is not small.

4.3.2 Two-stage Cluster Sampling

In section 2.3, Case 1, we considered the case of two-stage cluster sampling with equal cluster sizes, M , and SRS without replacement in both stages. The H-T estimator of the total Y is given by $\hat{Y} = (K/k) \sum_{i=1}^k \bar{y}_i$, where $\bar{y}_i = M \bar{y}_i$ and \bar{y}_i is the sample mean of the i -th sample cluster. The variance estimator of \hat{Y} is given by

$$\hat{V} = N^2 \left\{ \frac{1}{k} \left(1 - \frac{k}{K} \right) s_{1y}^2 + \frac{k}{K} \left(1 - \frac{m}{M} \right) \frac{1}{km} s_{2y}^2 \right\}, \quad (4.8)$$

where $s_{1y}^2 = \sum_{i=1}^k (\bar{y}_i - \bar{y})^2 / (k-1)$, $s_{2y}^2 = \sum_{i=1}^k s_{2i}^2 / k$ with s_{2i}^2 denoting the sample variance in the i -th cluster, \bar{y}_i is the i -th cluster sample mean and $\bar{y} = \sum_{i=1}^k \bar{y}_i / k$ is the overall sample mean (see Cochran 1977, pages 276-278).

For inverse sampling, we proposed approximate matching by selecting one element at random from the m sample elements in each sample cluster $i(i=1, \dots, k)$. Denote the values of the elements by y_1', \dots, y_k' . The inverse-sampling estimator of the total is given by

$\hat{Y}_g = g^{-1} \sum_{j=1}^g \hat{Y}_j^*$, where $\hat{Y}_j^* = (N/k) \sum_{i \in s_j} y_i'$. The inverse-sampling variance estimator, \hat{V}_g , is given by (3.4).

It is easy to verify that $\hat{Y}_\infty = \hat{Y}$ so that approximate matching preserves the original estimator \hat{Y} in the limit. On the other hand, it can be shown that \hat{V}_g tends to

$$\hat{V}_\infty = N^2 \frac{1}{k} s_{1y}^2 \quad (4.9)$$

as $g \rightarrow \infty$. It follows from (4.8) and (4.9) that

$$\begin{aligned} \frac{\hat{V}}{\hat{V}_\infty} &= 1 - \frac{k}{K} \left[1 - \left(1 - \frac{m}{M} \right) \frac{1}{km} \frac{s_{2y}^2}{s_{1y}^2} \right] \\ &\approx 1 - \frac{k}{K}, \end{aligned} \quad (4.10)$$

because the neglected term in (4.10) is of order $(mK)^{-1}$. It follows that \hat{V}_∞ again leads to overestimation of the variance if the sampling fraction k/K is not small.

5. COMBINED ESTIMATING EQUATIONS APPROACH

In this section, we study an estimating equations approach to inverse sampling. This approach permits valid inferences on nonlinear parameters such as ratios and “census” linear regression and logistic regression parameters. As noted in section 3, the inverse-sampling estimator $\hat{\theta}_g$, given by (3.1), has exactly the same bias as the subsample estimator $\hat{\theta}_1^*$, and the bias of $\hat{\theta}_1^*$ is of order m^{-1} , where m is the subsample size. As a result, the bias of $\hat{\theta}_g$ can be appreciable because m is usually very much smaller than the original sample size n . In fact, m could be as small as 2 for stratified two-stage cluster sampling designs with two sample clusters in each stratum. Moreover, for logistic regression and other cases, the calculation of $\hat{\theta}_j^*$ and $\hat{\theta}$ involves iterative solutions. As a result, the implementation of $\hat{\theta}_g$, and the inverse-sampling variance estimator \hat{V}_g , given by (3.3), could become computationally very cumbersome when the number of inverse-samples, g , is large. We avoid these difficulties using a combined estimating equations (CEE) approach.

In section 5.1, we consider the special case of a ratio of totals, $R = Y/X$, and spell out the “combined approach” suggested by HOS towards the end of section 3.1 of their paper. Section 5.2 gives the general theory and discusses special cases. The results of section 5.2 are applied in section 5.3 to a cluster correlated data set reported in Battese, Harter and Fuller (1988).

5.1 Ratio of Totals

HOS suggested a “combined approach” to estimate the ratio, R , of totals Y and X . We now explain this approach and relate it to the CEE approach in section 5.2.

Denote the estimator of R based on the j -th inverse-sample as $\hat{R}_j^* = \hat{Y}_j^*/\hat{X}_j^*$. The separate inverse-sampling estimator of R is then given by $\hat{R}_g = g^{-1} \sum_{j=1}^g \hat{R}_j^*$. HOS noted that the bias of \hat{R}_g can be large when the subsample size is small. They proposed to estimate the numerator and denominator of R separately, using the g subsamples. This leads to the “combined” inverse-sample estimator

$$\hat{R}_{gc} = \frac{\hat{Y}_g}{\hat{X}_g}, \quad (5.1)$$

where $\hat{Y}_g = g^{-1} \sum_{j=1}^g \hat{Y}_j^*$ and $\hat{X}_g = g^{-1} \sum_{j=1}^g \hat{X}_j^*$. Now, assuming that the final size of the “combined” sample is sufficiently large, it follows from (5.1) that

$$E(\hat{R}_{gc}) \approx \frac{E(\hat{Y}_g)}{E(\hat{X}_g)} = \frac{Y}{X} = R$$

under the conditions of Theorem 3. That is, \hat{R}_{gc} is approximately unbiased for R , regardless of the subsample size, provided g is sufficiently large.

Similarly, using the Taylor linearization approximation, we obtain the variance of \hat{R}_{gc} as

$$V(\hat{R}_{gc}) \approx \frac{1}{X^2} V(\tilde{U}_g), \quad (5.2)$$

where $\tilde{U}_g = g^{-1} \sum_{j=1}^g \tilde{U}_j^*$ is the inverse-sampling estimator of the total U of the residuals $u_i = y_i - Rx_i$, $i = 1, \dots, N$. Noting that \tilde{U}_g is the inverse-sampling estimator of a total, it follows from (3.3) that an inverse-sampling estimator of $V(\tilde{U}_g)$ is given by

$$\tilde{V}_{gU} = \frac{1}{g} \sum_{j=1}^g \tilde{V}_{jU} - \frac{1}{g} \sum_{j=1}^g (\tilde{U}_j^* - \tilde{U}_g)^2, \quad (5.3)$$

where \tilde{V}_{jU}^* is the variance estimator produced from the j -th subsample. Since R is unknown, we replace R by \hat{R}_{gc} in (5.3) to get the variance estimator \hat{V}_{gU} . Now, replacing X by its estimator \hat{X}_g and $V(\tilde{U}_g)$ by \tilde{V}_{gU} in (5.2), we get the inverse-sampling linearization variance estimator of \hat{R}_{gc} as

$$\hat{V}_L(\hat{R}_{gc}) = \frac{1}{\hat{X}_g^2} \hat{V}_{gU}. \quad (5.4)$$

Under the conditions of Theorem 4, $\hat{V}_L(\hat{R}_{gc})$ converges to the customary linearization variance estimator of the full-sample estimator $\hat{R} = \hat{Y}/\hat{X}$.

5.2 Nonlinear Parameters

(i) Full-sample estimating equations

A finite population parameter vector θ_N may be regarded as the solution to “census” estimating equations (EE’s):

$$\mathbf{U}(\boldsymbol{\theta}) = \sum_{k \in U} \mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{0}, \quad (5.5)$$

where $\sum_{k \in U}$ denotes the summation over the finite population U of size N , and the estimating functions $\mathbf{u}_k(\boldsymbol{\theta})$ are suitably chosen (Binder 1983; Godambe and Thompson 1986). For example, consider the scalar case of (5.5) and let $u_k(\boldsymbol{\theta}) = y_k - \theta$ in (5.5). This gives the population mean $\theta_N = Y$. Similarly, letting $u_k(\boldsymbol{\theta}) = y_k - \theta x_k$ we get the ratio of totals: $\theta_N = R = Y/X$. The choice $\mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{x}_k(y_k - \mu_k(\boldsymbol{\theta}))$ with $\mu_k(\boldsymbol{\theta}) = \mathbf{x}_k^T \boldsymbol{\theta}$ gives the census linear regression parameters

$$\boldsymbol{\theta}_N = \left(\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \sum_{k \in U} \mathbf{x}_k y_k.$$

The choice $\mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{x}_k(y_k - \mu_k(\boldsymbol{\theta}))$ with $\log[\mu_k(\boldsymbol{\theta})/(1 - \mu_k(\boldsymbol{\theta}))] = \mathbf{x}_k^T \boldsymbol{\theta}$ gives the census logistic regression parameters $\boldsymbol{\theta}_N$. Kovacevic and Binder (1997) give estimating functions, $\mathbf{u}_k(\boldsymbol{\theta})$, that lead to various measures of income inequality, such as the Gini index and the polarization index.

The full-sample estimating equations are given by

$$\hat{\mathbf{U}}(\boldsymbol{\theta}) = \sum_{k \in s_0} w_k \mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{0}, \quad (5.6)$$

where w_k is the survey weight attached to $k \in s_0$; in particular, $w_k = 1/\pi_k$ if the H-T estimator of $\mathbf{U}(\boldsymbol{\theta})$ is used. The solution to (5.6) gives the full-sample estimator $\hat{\boldsymbol{\theta}}$ which, in general, is nonlinear and hence biased. We assume that the size of the original sample, s_0 , is large enough to neglect the bias of $\hat{\boldsymbol{\theta}}$. For logistic regression and other complex cases, it is necessary to solve (5.6) iteratively to obtain the full-sample estimator $\hat{\boldsymbol{\theta}}$. The Newton-Raphson (N-R) algorithm is commonly used to solve (5.6). The r -th step of the N-R algorithm is given by

$$\hat{\boldsymbol{\theta}}^{(r)} = \hat{\boldsymbol{\theta}}^{(r-1)} + \hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}}^{(r-1)}) \hat{\mathbf{U}}(\hat{\boldsymbol{\theta}}^{(r-1)}), \quad (5.7)$$

where $\hat{\boldsymbol{\theta}}^{(r-1)}$ is the value of $\hat{\boldsymbol{\theta}}$ obtained at the $(r-1)$ -th iteration, and $\hat{\mathbf{U}}(\hat{\boldsymbol{\theta}}^{(r-1)})$ and $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}^{(r-1)})$ are the values of $\hat{\mathbf{U}}(\boldsymbol{\theta})$ and $\hat{\mathbf{J}}(\boldsymbol{\theta}) = -\partial \hat{\mathbf{U}}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^T = -\sum_{k \in s_0} w_k \partial \mathbf{u}_k(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^T$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(r-1)}$. Iterating the N-R algorithm to convergence produces the estimator $\hat{\boldsymbol{\theta}}$ as well as the observed information matrix $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})$.

Under regularity conditions, Binder (1983) obtained a Taylor linearization estimator of the covariance matrix, $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$, of $\hat{\boldsymbol{\theta}}$ as

$$\hat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})]^{-1} \hat{\mathbf{V}}[\hat{\mathbf{U}}(\hat{\boldsymbol{\theta}})] [\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})]^{-1}, \quad (5.8)$$

where $\hat{\mathbf{V}}[\hat{\mathbf{U}}(\hat{\boldsymbol{\theta}})]$ is a variance estimator of the estimated total, $\hat{\mathbf{U}}(\boldsymbol{\theta})$, of the $\mathbf{u}_k(\boldsymbol{\theta})$'s evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. For example, if $u_k(\boldsymbol{\theta}) = y_k - \theta x_k$ then $\hat{\boldsymbol{\theta}} = \sum_{k \in s_0} w_k y_k / \sum_{k \in s_0} w_k x_k = \hat{Y}/\hat{X} = \hat{R}$ is the ratio estimator, and (5.8) reduces to the customary linearization variance estimator

$$\hat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}) = \frac{1}{\hat{X}^2} \hat{\mathbf{V}} \left[\sum_{k \in s_0} w_k u_k(\hat{\boldsymbol{\theta}}) \right], \quad (5.9)$$

noting that $\hat{J}(\boldsymbol{\theta}) = \sum_{k \in s_0} w_k x_k = \hat{X}$.

(ii) Separate estimating equations

The separate inverse-sampling estimators $\hat{\boldsymbol{\theta}}_j^*$, $j = 1, \dots, g$ are obtained by solving the separate estimating equations (SEE)

$$\hat{\mathbf{U}}_j^*(\boldsymbol{\theta}) = \frac{N}{m} \sum_{k \in s_j^*} \mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{0}; j = 1, \dots, g. \quad (5.10)$$

In general, we require g iterative solutions to get $\hat{\boldsymbol{\theta}}_1^*, \dots, \hat{\boldsymbol{\theta}}_g^*$. The inverse-sampling estimator of $\boldsymbol{\theta}$ is then given by

$$\hat{\boldsymbol{\theta}}_g^* = \frac{1}{g} \sum_{j=1}^g \hat{\boldsymbol{\theta}}_j^*. \quad (5.11)$$

It follows from (5.11) that $\hat{\boldsymbol{\theta}}_\infty^* = E(\hat{\boldsymbol{\theta}}_1^* | s_0)$ and $E(\hat{\boldsymbol{\theta}}_\infty^*) = E(\hat{\boldsymbol{\theta}}_1^*)$. Assuming first moment matching with SRS, it follows from (5.10) that the bias $E(\hat{\boldsymbol{\theta}}_1^*) - \boldsymbol{\theta}$ is of order m^{-1} , where m is the subsample size. The inverse-sampling estimator of $\mathbf{V}(\hat{\boldsymbol{\theta}}_g^*)$ is given by

$$\hat{\mathbf{V}}_g^* = \frac{1}{g} \sum_{j=1}^g \hat{\mathbf{V}}_j^* - \frac{1}{g} \sum_{j=1}^g (\hat{\boldsymbol{\theta}}_j^* - \hat{\boldsymbol{\theta}}_g^*) (\hat{\boldsymbol{\theta}}_j^* - \hat{\boldsymbol{\theta}}_g^*)^T, \quad (5.12)$$

where $\hat{\mathbf{V}}_j^*$ is given by

$$\hat{\mathbf{V}}_j^* = [\hat{\mathbf{J}}_j^*(\hat{\boldsymbol{\theta}}_j^*)]^{-1} \hat{\mathbf{V}}[\hat{\mathbf{U}}_j^*(\hat{\boldsymbol{\theta}}_j^*)] [\hat{\mathbf{J}}_j^*(\hat{\boldsymbol{\theta}}_j^*)]^{-1}, \quad (5.13)$$

$\hat{\mathbf{V}}[\hat{\mathbf{U}}_j^*(\hat{\boldsymbol{\theta}}_j^*)]$ is the variance estimator of the j -th subsample total $\hat{\mathbf{U}}_j^*(\boldsymbol{\theta})$, denoted $\hat{\mathbf{V}}_{jU}^*$ (see equation (5.19) below), evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_j^*$, and $\hat{\mathbf{J}}_j^*(\hat{\boldsymbol{\theta}}_j^*)$ is $\hat{\mathbf{J}}_j^*(\boldsymbol{\theta}) = -\partial \hat{\mathbf{U}}_j^*(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^T$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_j^*$.

(iii) Combined estimating equations

We now obtain a combined estimating equations (CEE) estimator $\hat{\boldsymbol{\theta}}_{gc}^*$ that leads to valid inference regardless of the subsample size m . We simply combine the g equations in (5.10) before solving for $\boldsymbol{\theta}$. This leads to combined estimating equations

$$\hat{\mathbf{U}}_{gc}^*(\boldsymbol{\theta}) = \frac{1}{g} \sum_{j=1}^g \hat{\mathbf{U}}_j^*(\boldsymbol{\theta}) = \frac{1}{g} \sum_{j=1}^g \frac{N}{m} \sum_{k \in s_j^*} \mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{0}. \quad (5.14)$$

In general, we solve (5.14) using the N-R iterations (5.7) with $\hat{\mathbf{U}}(\hat{\boldsymbol{\theta}}^{(r-1)})$ changed to $\hat{\mathbf{U}}_{gc}(\hat{\boldsymbol{\theta}}^{(r-1)})$ and $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}^{(r-1)})$ changed to $\hat{\mathbf{J}}_{gc}(\hat{\boldsymbol{\theta}}^{(r-1)})$, where

$$\hat{\mathbf{J}}_{gc}(\boldsymbol{\theta}) = -\frac{\partial \hat{\mathbf{U}}_{gc}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = -\frac{1}{g} \sum_{j=1}^g \frac{N}{m} \sum_{k \in s_j^*} \frac{\partial \mathbf{u}_k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}. \quad (5.15)$$

At convergence, we obtain the CEE estimator $\hat{\boldsymbol{\theta}}_{gc}^*$ as well as the observed information matrix $\hat{\mathbf{J}}_{gc}(\hat{\boldsymbol{\theta}}_{gc}^*)$. Note that we solve the combined estimating equations (5.14) only once to get $\hat{\boldsymbol{\theta}}_{gc}^*$, unlike the separate estimating equations method

that solves the g equations (5.10) to get $\hat{\theta}_1^*, \dots, \hat{\theta}_g^*$ and $\hat{\theta}_g^* = \sum_{j=1}^g \hat{\theta}_j^* / g$.

To illustrate the proposed CEE method, consider the special case of ratio $\theta_N = R$, in which case $u_k(\theta) = y_k - \theta x_k$. The combined estimating equations (5.14) reduce to $\hat{Y}_g - \theta \hat{X}_g = 0$ and the solution $\hat{\theta}_{gc}$ is identical to the combined inverse-sampling estimator \hat{R}_{gc} given by (5.1).

Assuming first moment matching with SRS, it follows from (5.14) that $\hat{\theta}_{\infty c}$ is a solution of

$$\hat{U}_{\infty c}(\theta) = E[\hat{U}_1^*(\theta) | s_0] = \hat{U}(\theta) = \mathbf{0}. \quad (5.16)$$

As a result, $\hat{\theta}_{\infty c} = \hat{\theta}$ regardless of the subsample size m . Thus, the bias of $\hat{\theta}_{gc}$ is of the same order as the bias of $\hat{\theta}$ for sufficiently large g , regardless of the subsample size, m .

We now apply Binder's (1983) method to $\hat{U}_{gc}(\theta)$ to get a linearization inverse-sampling estimator of $\mathbf{V}(\hat{\theta}_{gc})$. It follows from (5.8) that

$$\hat{\mathbf{V}}_L(\hat{\theta}_{gc}) = [\hat{\mathbf{J}}_{gc}(\hat{\theta}_{gc})]^{-1} \hat{\mathbf{V}}[\hat{U}_{gc}(\hat{\theta}_{gc})][\hat{\mathbf{J}}_{gc}(\hat{\theta}_{gc})]^{-1}, \quad (5.17)$$

where $\hat{\mathbf{V}}[\hat{U}_{gc}(\hat{\theta}_{gc})]$ is the variance estimator of the estimated total, $\hat{U}_{gc}(\theta)$, of the $\mathbf{u}_k(\theta)$'s evaluated at $\hat{\theta} = \hat{\theta}_{gc}$. Note that $\hat{\mathbf{J}}_{gc}(\hat{\theta}_{gc})$ is obtained at the convergence of the N-R algorithm applied to (5.14).

Since $\hat{U}_{gc}(\theta)$ is the inverse-sampling estimator of the total $\mathbf{U}(\theta)$, it follows that the inverse-sampling estimator of $\mathbf{V}[\hat{U}_{gc}(\theta)]$ is given by

$$\begin{aligned} \tilde{\mathbf{V}}_{gU} &= \frac{1}{g} \sum_{j=1}^g \tilde{\mathbf{V}}_{jU}^* \\ &\quad - \frac{1}{g} \sum_{j=1}^g [\hat{U}_j^*(\theta) - \hat{U}_{gc}(\theta)][\hat{U}_j^*(\theta) - \hat{U}_{gc}(\theta)]^T, \end{aligned} \quad (5.18)$$

where $\tilde{\mathbf{V}}_{jU}^*$ is the SRS variance estimator from the j -th subsample, assuming second moment matching. If the matching is with respect to SRS without replacement, then

$$\begin{aligned} \tilde{\mathbf{V}}_{jU}^* &= \frac{N^2}{m} \left(1 - \frac{m}{N}\right) \frac{1}{m-1} \\ &\quad \times \sum_{k \in s_j^*} \left[\mathbf{u}_k(\theta) - \frac{1}{m} \sum_{k \in s_j^*} \mathbf{u}_k(\theta) \right] \\ &\quad \left[\mathbf{u}_k(\theta) - \frac{1}{m} \sum_{k \in s_j^*} \mathbf{u}_k(\theta) \right]^T \end{aligned} \quad (5.19)$$

In the case of matching to SRS with replacement, we replace $1 - m/N$ by 1 in (5.19). Now substituting $\hat{\theta}_{gc}$ for θ in (5.18) we get

$$\begin{aligned} \hat{\mathbf{V}}[\hat{U}_{gc}(\hat{\theta}_{gc})] &= \frac{1}{g} \sum_{j=1}^g \hat{\mathbf{V}}_{jU}^* \\ &\quad - \frac{1}{g} \sum_{j=1}^g \hat{U}_j^*(\hat{\theta}_{gc}) \hat{U}_j^*(\hat{\theta}_{gc})^T = \hat{\mathbf{V}}_{gU}, \end{aligned} \quad (5.20)$$

where $\hat{\mathbf{V}}_{jU}^*$ is obtained from (5.19) by substituting $\hat{\theta}_{gc}$ for θ . Note that $\hat{U}_{gc}(\hat{\theta}_{gc}) = \mathbf{0}$.

Under second moment matching with SRS, as $g \rightarrow \infty$, it is easy to verify that $\hat{\mathbf{V}}_L(\hat{\theta}_{gc})$ converges to Binder's estimator $\hat{\mathbf{V}}_L(\hat{\theta})$ given by (5.8). This follows by noting that $\hat{\theta}_{\infty c} = \hat{\theta}$, $\hat{\mathbf{J}}_{\infty c}(\theta) = \hat{\mathbf{J}}(\theta)$ and $\mathbf{V}_{\infty c} = \hat{\mathbf{V}}[\hat{U}(\theta)]$ under second moment matching with SRS. Thus, the covariance estimator $\hat{\mathbf{V}}_L(\hat{\theta}_{gc})$ provides valid inferences on θ for large number of subsamples, g , regardless of the subsample size, m .

To illustrate the calculation of the linearization inverse-sampling estimator $\hat{\mathbf{V}}_L(\hat{\theta}_{gc})$, given by (5.17), consider the special case of a ratio $\theta_N = R$ with $u_k(\theta) = y_k - \theta x_k$. We have

$$\tilde{\mathbf{V}}_{jU}^* = \frac{N^2}{m} \left(1 - \frac{m}{N}\right) \frac{1}{m-1} \sum_{k \in s_j^*} [u_k(\theta) - \bar{u}_j^*(\theta)]^2, \quad (5.21)$$

where $\bar{u}_j^*(\theta) = \bar{y}_j - \theta \bar{x}_j^*$ and $(\bar{y}_j^*, \bar{x}_j^*)$ are the j -th subsample means. Further,

$$\hat{\mathbf{J}}_{gc}(\theta) = \frac{N}{g} \sum_{j=1}^g \bar{x}_j^* = \hat{X}_g \quad (5.22)$$

and

$$\hat{U}_j^*(\theta) = N(\bar{y}_j^* - \theta \bar{x}_j^*). \quad (5.23)$$

It now follows from (5.21) – (5.23) that the CEE-based linearization estimator (5.17) is identical to the inverse-sampling linearization variance estimator (5.4).

Turning to linear regression with $\mathbf{u}_k(\theta) = \mathbf{x}_k(y_k - \mathbf{x}_k^T \theta)$, we have

$$\begin{aligned} \tilde{\mathbf{V}}_{jU}^* &= \frac{N^2}{m} \left(1 - \frac{m}{N}\right) \frac{1}{m-1} \\ &\quad \times \sum_{k \in s_j^*} [\mathbf{u}_k(\theta) - \bar{\mathbf{u}}_j^*(\theta)][\mathbf{u}_k(\theta) - \bar{\mathbf{u}}_j^*(\theta)]^T, \end{aligned} \quad (5.24)$$

where $\bar{\mathbf{u}}_j^*(\theta) = m^{-1} \sum_{k \in s_j^*} \mathbf{u}_k(\theta)$. Also,

$$\hat{\mathbf{J}}_{gc}(\theta) = \frac{N}{g} \sum_{j=1}^g \frac{1}{m} \sum_{k \in s_j^*} \mathbf{x}_k \mathbf{x}_k^T$$

and

$$\hat{U}_j^*(\theta) = \frac{N}{m} \sum_{k \in s_j^*} \mathbf{x}_k (y_k - \mathbf{x}_k^T \theta).$$

Finally, consider the case of logistic regression with $\mathbf{u}_k(\theta) = \mathbf{x}_k(y_k - \mu_k(\theta))$. In this case, $\tilde{\mathbf{V}}_{jU}^*$ is given by (5.24) with $\mathbf{u}_k(\theta) = \mathbf{x}_k(y_k - \mu_k(\theta))$. Also,

$$\hat{\mathbf{J}}_{gc}(\theta) = \frac{N}{g} \sum_{j=1}^g \frac{1}{m} \sum_{k \in s_j^*} \mu_k(\theta)(1 - \mu_k(\theta)) \mathbf{x}_k \mathbf{x}_k^T,$$

and

$$\hat{\mathbf{U}}_j^*(\boldsymbol{\theta}) = \frac{N}{m} \sum_{k \in s_j^*} \mathbf{x}_k (y_k - \mu_k(\boldsymbol{\theta})).$$

It is important to note again that the estimator $\hat{\boldsymbol{\theta}}_{gc}$ and the associated covariance estimator $\hat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}_{gc})$ can be implemented from a microdata with data from g subsamples, each of size m . Neither the survey weights w_k nor the cluster identifiers are needed so that confidentiality of microdata may be preserved.

5.3 An Example

We now use a data set reported in Battese, Harter and Fuller (1988) to illustrate how the separate and combined estimating equations methods perform. The data were collected from $k = 12$ counties in north-central Iowa. The counties were divided into area segments and a sample of area segments was selected from each county. Here counties represent clusters and sample area segments within a county represent elements. The number of sample area segments (m_i) ranged from 1 to 5 giving a total of $n = \sum_{i=1}^k m_i = 37$ sample elements. For each sample element (i, j), Battese *et al.* (1988) gave the number of reported hectares of corn (y_{ij}) obtained by interviewing farm operators and the number of pixels classified as corn (x_{1ij}) and soybeans (x_{2ij}) obtained from remote sensing satellite readings ($j = 1, \dots, m_i; i = 1, \dots, k$). Data from one of the sample area segments were suspected to be erroneous and hence excluded from the analysis. Thus we have $n = 36$ observations (y_{ij}, x_{ij}).

For illustration, we treat the sample as if it was selected by the following two stage cluster sampling: (i) In the first stage, counties were selected with replacement and with probabilities proportional to the number of area segments

M_i in the counties. (ii) In the second stage, sample area segments were selected by simple random sampling without replacement from each selected county. We consider two parameters: (i) population ratio $\theta = R = Y/X$, where Y and X are the population totals of y and x ; (ii) census regression coefficient of y on x , $\boldsymbol{\theta} = \mathbf{B} = (\sum_{l \in U} \mathbf{x}_l \mathbf{x}_l^T)^{-1} (\sum_{l \in U} \mathbf{x}_l y_l)$, where $\mathbf{x}_l = (1, x_{1l}, x_{2l})^T$ and l denotes a population element.

For selected values of g , we generated g inverse-samples independently using the procedure for Case 2 in section 2.3. We then used the g subsamples to estimate R using the separate estimating equations (SEE) method and the combined estimating equations (CEE) method given in section 5. The corresponding variance estimates and the linearization variance estimates of the full-sample estimates $\hat{\boldsymbol{\theta}}$ were computed.

Table 1 reports the full-sample estimate \hat{R} , the SEE estimate \hat{R}_g , the CEE estimate \hat{R}_{gc} and the corresponding variance estimates. It is clear from Table 1 that both CEE and SEE perform well in tracking the full-sample estimate \hat{R} and the corresponding linearization full-sample variance estimate even for $g = 500$.

Table 2 reports the results for the regression coefficients $\mathbf{B} = (B_0, B_1, B_2)^T$. As g increases, both SEE and CEE seem to track the full-sample estimates \hat{B}_1 and \hat{B}_2 , while SEE leads to slightly larger value for \hat{B}_0 . However, the SEE variance estimates perform poorly, even for very large $g = 10,000$ in tracking the linearization full-sample variance estimates, with SEE value about one-half of the corresponding full-sample value for B_0 and B_1 . On the other hand, the CEE variance estimates perform very well in tracking the full-sample variance estimates, confirming the theory.

Table 1
Estimation of Population Ratio R

	$g = 500$			$g = 1,000$		$g = 5,000$	
	Full-sample	CEE	SEE	CEE	SEE	CEE	SEE
Estimate	0.4096	0.4101	0.41	0.4096	0.4095	0.4095	0.4094
Variance Estimate $\times 10^{-4}$	1.9513	1.8769	1.8508	1.8482	1.8302	1.932	1.9178

Table 2
Estimation of Census Regression Parameters, B_0, B_1 and B_2

	$g = 500$			$g = 1,000$		$g = 10,000$	
	Full-sample	CEE	SEE	CEE	SEE	CEE	SEE
Est. of B_0	53.3588	49.9532	52.6649	53.5876	56.7143	53.2401	56.3196
Est. of B_1	0.3176	0.3251	0.318	0.3171	0.3086	0.3179	0.31
Est. of B_2	-0.1326	-0.1258	-0.1302	-0.133	-0.1378	-0.1324	-0.1377
B_0 : Var. Est.	416.1609	457.5178	293.8789	407.3107	224.0846	437.961	251.395
B_1 : Var. Est. $\times 10^{-3}$	2.1153	2.2925	1.164	1.9127	0.5354	2.2366	0.8882
B_2 : Var. Est. $\times 10^{-3}$	2.7369	3.0352	2.4811	2.7226	2.3174	2.8028	2.3229

6. CONCLUDING REMARKS

In this paper we have presented some theory of inverse sampling. Efficiency of inverse sampling is increased by drawing repeated subsamples and then combining the results from the subsamples.

For estimating a total, we obtained conditions for the limiting inverse-sampling estimator to approach the full-sample estimator (Theorem 3) and for the limiting inverse-sampling variance estimator to approach the full-sample variance estimator (Theorem 4). For estimating complex parameters, we proposed a combined estimating equations (CEE) approach and demonstrated its advantages over separate estimating equations (SEE) approach (section 5).

We have studied inverse sampling algorithms for some sampling designs in section 2. But further work is needed to cover other sampling designs and also to avoid the limitations noted in section 2.

We are studying various extensions to include post-stratified full-sample estimators, analysis of categorical survey data, clustered survival data (Binder 1992) and longitudinal survey data.

ACKNOWLEDGEMENTS

The authors wish to thank the Associate Editor and the referees for constructive suggestions. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

APPENDIX

Proofs of Theorems

Proof of theorem 1

Result 1 follows directly from (3.1) on noting that conditional on s_0 , $\hat{\theta}_1^*, \dots, \hat{\theta}_g^*$ are independent identically distributed (i.i.d.) bounded random variables.

Result 2 follows from the standard relationship between conditional and unconditional expectations:

$$E(\hat{\theta}_g) = E[E(\hat{\theta}_g | s_0)] = E[E(\hat{\theta}_1^* | s_0)] = E(\hat{\theta}_1^*).$$

Result 3 follows from the corresponding result for variances, and the conditional independence of the $\hat{\theta}_j^*$'s given s_0 :

$$\begin{aligned} \text{Var}(\hat{\theta}_g) &= \text{Var}[E(\hat{\theta}_g | s_0)] + E[\text{Var}(\hat{\theta}_g | s_0)] \\ &= \text{Var}(\hat{\theta}_\infty) + \frac{1}{g} E[\text{Var}(\hat{\theta}_1^* | s_0)]. \end{aligned}$$

Result 4 follows directly from Result 3.

Proof of theorem 2

Theorem 2 follows from applying Results 3 of Theorem 1 with $g = 1$ to obtain

$$\text{Var}(\hat{\theta}_\infty^*) = \text{Var}(\hat{\theta}_1^*) - E[\text{Var}(\hat{\theta}_1^* | s_0)],$$

and then substituting this expression for $\text{Var}(\hat{\theta}_\infty)$ in Result 3 of Theorem 1 for general g .

Proof of theorem 3

We have

$$\hat{Y}_j^* = \sum_{i \in s_j^*} \frac{y_i}{\pi_i^*} = \sum_{i \in s_0} \frac{y_i I_{ij}^*(s_0)}{\pi_i^*},$$

where $I_{ij}^*(s_0)$ takes the value 1 if the i -th unit is included in the j -th subsample s_j^* and 0 otherwise, and π_i^* is the corresponding (unconditional) inclusion probability. Thus

$$\hat{Y}_\infty^* = E[\hat{Y}_1^* | s_0] = \sum_{i \in s_0} \frac{y_i \tilde{\pi}_i(s_0)}{\pi_i^*}.$$

This is equal to $\hat{Y} = \sum_{i \in s_0} (y_i / \pi_i)$, the H-T estimator for the original design, if and only if $\tilde{\pi}_i(s_0) = \pi_i^* / \pi_i$.

Proof of theorem 4

Conditional on s_0 , it follows from (3.3) that $\hat{V}_{g, \text{HT}}$ converges almost surely to

$$\hat{V}_{\infty, \text{HT}} = E(\hat{V}_{1, \text{HT}}^* | s_0) - \text{Var}(\hat{Y}_1^* | s_0) \quad (\text{A.1})$$

as $g \rightarrow \infty$. Now, noting that $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il} = \pi_{il}^* / \pi_{il}$, we get

$$\begin{aligned} E(\hat{V}_{1, \text{HT}}^* | s_0) &= \sum_{i, l \in s_0} \sum_{i, l \in s_0} \frac{\pi_{il}^* - \pi_i^* \pi_l^*}{\pi_i^* \pi_l^* \pi_{il}} \tilde{\pi}_{il} y_i y_l \\ &= \sum_{i, l \in s_0} \sum_{i, l \in s_0} \left(\frac{\tilde{\pi}_{il}}{\pi_i^* \pi_l^*} - \frac{1}{\pi_{il}} \right) y_i y_l \end{aligned} \quad (\text{A.2})$$

Further,

$$\begin{aligned} \text{Var}(\hat{Y}_1^* | s_0) &= \sum_{i, l \in s_0} \sum_{i, l \in s_0} (\tilde{\pi}_{il} - \tilde{\pi}_i \tilde{\pi}_l) \frac{y_i}{\pi_i^*} \frac{y_l}{\pi_l^*} \\ &= \sum_{i, l \in s_0} \sum_{i, l \in s_0} \left(\frac{\tilde{\pi}_{il}}{\pi_i^* \pi_l^*} - \frac{1}{\pi_i \pi_l} \right) y_i y_l. \end{aligned} \quad (\text{A.3})$$

It now follows from (A.1) - (A.3) that $\hat{V}_{\infty, \text{HT}} = \hat{V}_{\text{HT}}^*$.

Proof of theorem 5

Conditional on s_0 , it follows from (3.3) that

$$\hat{V}_{\infty, \text{SYG}} = E(\hat{V}_{1, \text{SYG}}^* | s_0) - \text{Var}(\hat{Y}_1^* | s_0) \quad (\text{A.4})$$

where

$$\text{Var}(\hat{Y}_1^* | s_0) = \sum_{i < l \in s_0} \sum_{i < l \in s_0} (\tilde{\pi}_i \tilde{\pi}_l - \tilde{\pi}_{il}) \left(\frac{y_i}{\pi_i^*} - \frac{y_l}{\pi_l^*} \right)^2, \quad (\text{A.5})$$

provided the subsample size is also fixed (Cochran 1977, page 260). Further,

$$E\left(\hat{V}_{1, \text{SYG}}^* \mid s_0\right) = \sum_{i < l \in s_0} \sum_{\pi_{il}} \frac{\left(\pi_i^* \pi_l^* - \pi_{il}^*\right)}{\pi_{il}} \left(\frac{y_i}{\pi_i^*} - \frac{y_l}{\pi_l^*}\right)^2. \quad (\text{A.6})$$

It now follows that

$$\hat{V}_{\infty, \text{SYG}} = \sum_{i < l \in s_0} \sum_{\pi_{il}} \frac{\pi_i \pi_l - \pi_{il}}{\pi_{il}} \tilde{\pi}_i \tilde{\pi}_l \left(\frac{y_i}{\pi_i \tilde{\pi}_i} - \frac{y_l}{\pi_l \tilde{\pi}_l}\right)^2, \quad (\text{A.7})$$

Comparing (A.7) and (A.4) we see that $\hat{V}_{\infty, \text{SYG}} \neq \hat{V}_{\text{SYG}}$.

REFERENCES

- BATTESE, G.E., HARTER, R.M. and FULLER W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*. 83, 28-36.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*. 51, 279-292.
- BINDER, D.A. (1992). Fitting Cox's proportional hazard models from survey data. *Biometrika*. 79, 139-147.
- COCHRAN, W.G. (1977). *Sampling Techniques*. Third Edition; New York: John Wiley & Sons, Inc.
- HINKINS, S., OH, H.L. and SCHEUREN, F. (1997). Inverse sampling design algorithms. *Survey Methodology*. 23, 11-21.
- HOFFMAN, E.B., SEN, P.K. and WEINBERG, C.R. (2001). Within-cluster resampling. *Biometrika*. 88, 1121-34.
- KOVACEVIC, M.S., and BINDER, D.A. (1997). Variance estimation for measures of income inequality and polarization – the estimating equations approach. *Journal of Official Statistics*. 13, 41-58.
- RAO., J.N.K., and SCOTT, A.J. (1992). A simple method for the analysis of clustered binary data. *Biometrics*. 48, 577-585.
- RAO., J.N.K., and SCOTT, A.J. (1999). A simple method for analysing overdispersion in clustered Poisson data. *Statistics in Medicine*. 18, 1373-1385.
- SKINNER, C.J., HOLT, D. and SMITH, T.M.F. (Eds.)(1989). *Analysis of Complex Surveys*. Chichester: Wiley.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Comment

JOHN L. ELTINGE¹

1. OVERVIEW

Rao, Scott and Benhin (henceforth RSB), in conjunction with Hinkins, Oh and Scheuren (1997) (henceforth HOS), have produced a fascinating set of ideas and methods for inverse sampling. This discussion will highlight several related ideas and practical issues that the survey community is likely to encounter as it considers practical applications of inverse sampling. Section 2 notes some relationships between standard probability weights and the random weights implicitly constructed through repeated inverse sampling. Section 3 discusses two types of approximations that may arise in variance estimation from inverse sample data. Section 4 considers the practical operational simplifications that may result from inverse sampling in some cases. Section 5 discusses the use of inverse sample data with standard (simple random sample-based) graphical methods. Section 6 explores the potential benefits and limitations of inverse sampling in attempts to reduce identification risk in public-use datasets.

2. POINT ESTIMATION: INVERSE SAMPLING AS A FORM OF FILTERING

Borrowing some ideas from the sampling, signal processing and confidentiality literature (*e.g.*, Duncan and Pearson 1991), we can think of a point estimator as the result of multiple steps of “filtering” of observations from a population. For example, in construction of a standard Horvitz-Thompson estimator of a population total, a set of population values can be viewed as passing through two stages of filters corresponding, respectively, to the selection of sample units and to the inverse-probability weighting of those units. Similarly, the point estimator (4.1) in RSB may be viewed as the result of two stages of filtering, where the second stage now corresponds to weighting by a random factor determined by the number of times a given sample unit appears in the g repeated inverse samples. Under conditions, the filter weights in (4.1) converge to the inverse-probability weights in the Horvitz-Thompson point estimator as g increases. In this sense, we can view the point estimator (4.1) as an approximation to the Horvitz-Thompson estimator. Similar comments apply to the general nonlinear point estimators and general inverse samples considered in RSB.

In addition, single inverse sampling can be viewed as a special type of two-phase sampling in which the second-phase selection rates are proportional to the inverses of the first-phase sampling rates. This leads naturally to the question of whether standard ideas from two-phase sampling can lead to efficiency gains in either single or multiple inverse sampling. For example, recall that in standard two-phase sampling, one can often improve efficiency by using ratio or regression-based adjustments in conjunction with auxiliary variables X observed for all first-phase sampling units. See *e.g.*, Särndal, Swensson and Wretman (1992, Chapter 9). Similarly here, one could construct a public-use dataset consisting of a single or multiple inverse sample dataset accompanied by estimated totals (based on the full complex sample) for a vector of auxiliary variables X . Also, some additional auxiliary information would be required for consistent variance estimation. Given sufficiently strong auxiliary variables X , the resulting ratio or regression-based adjusted point estimators could help to improve the precision of inverse-sample-based analyses. This in turn could reduce the number of inverse subsamples required to ensure that the regression-adjusted multiple-inverse-sample point estimator has a variance that is sufficiently small.

More generally, in many complex-survey cases (outside of two-phase designs), standard weighted point estimators also go beyond direct use of inverse-probability weights to incorporate auxiliary information through, *e.g.*, ratio or regression adjustments. Also, in some cases, one reduces the numerical values of certain extreme probability weights, in an attempt to avoid problems with variance inflation induced by influential observations. See, *e.g.*, Zaslavsky, Schenker and Belin (2001). A natural question is whether one could modify the inverse sampling algorithm so that the inverse design is “tuned” to the adjusted weights rather than the direct inverse-probability weights. This would be of serious interest for cases in which adjusted-weight point estimators are expected to have a substantially smaller mean squared error than inverse-probability-weight point estimators. For cases in which this modified approach is advisable, it would be of interest to study corresponding ways in which to extend the RSB approach to variance estimation.

¹ John L. Eltinge, Office of Survey Methods Research, U.S. Bureau of Labor Statistics. E-mail: Eltinge_J@bls.gov.

3. APPROXIMATIONS EMPLOYED IN VARIANCE ESTIMATION AND INFERENCE

For some complex designs, HOS and RSB noted that exact extraction of a simple random sample may be impossible, or may lead to a very small inverse sample, which in turn requires compensation through the use of a very large g . Consequently, sections 2 and 4.3 of RSB consider approximate matching methods, and section 4.1 considers inverse sampling that may produce a design that is simpler than the original complex design, but is more complex than a simple random design.

In parallel with this, recall that some of the sampling literature considers variance estimators that are based on approximations to the true sample design. One example is variance estimation based on stratum collapse. See, *e.g.*, Rust and Kalton (1987) and references cited therein. In addition, Korn and Graubard (1995, sections 4.2 and 4.3) consider variance estimators that ignore the original primary-sample-unit-level clustering and treat secondary sample units as if they were primary sample units.

In some cases, these approaches may be problematic, while in other cases they may produce satisfactory variance estimators. For the latter cases, one could consider development of an inverse sample procedure based on the approximate "variance estimation design" rather than on the true sample design. Under that approach, it would be of special interest to consider the relative magnitudes of errors associated with, respectively, sampling under the original design, the approximation error in the "variance estimation design," and the additional error induced through use of a finite number of inverse samples.

4. OPERATIONAL SIMPLICITY

In principle, most point estimation, variance estimation and inference methods that have been developed for simple random sample data can be extended to handle complex sample data. However, the work required for such extensions is often nontrivial, and may discourage many potential analysts from making efficient use of the available data. In an informal sense, data analysts often appear to choose their analytic approaches based on a rough cost-benefit evaluation, in which they will focus on analyses that they believe will offer them most or all of the scientific insights available from the data, while not requiring an investment in analytic effort that they consider disproportionate to the potential scientific benefit. Statisticians and subject-area data analysts may often have different views regarding the relative costs and scientific benefits of a given analytic effort. In some cases, inverse sampling may help to ameliorate the effects of these differing views.

In particular, as indicated by RSB and HOS, an investment by a statistical agency in construction of inverse samples may lead to some reduction in the burden

encountered by a given analyst. This investment may be especially worthwhile if both of the following conditions are satisfied.

- (a) An analyst intends to carry out a large number of different analyses on a single survey dataset; lacks appropriate complex-survey software for many (or all) of the intended analyses; and perceives the programming of complex-survey procedures to require a major investment of effort.
- (b) The additional computational steps required for point estimation (*e.g.*, the averaging carried out in the point estimators (3.1) or (4.1), or the combined estimating equation (5.14)) or variance estimation (*e.g.*, the variance estimators (3.3), (3.4), (5.18) or (5.20)) impose a relatively low incremental burden on the analyst, or can be absorbed into the analytic software in a form that is transparent to the analyst.

5. GRAPHICAL DISPLAYS

Hinkins *et al.* (1997, page 19) and Scheuren (1997) have noted the potential for application of inverse sampling to statistical graphics for complex survey data. For example, Scheuren (1997) noted that many methods of statistical graphics (*e.g.*, scatterplots) have been developed primarily for sets of independent and identically distributed observations. Direct application of these methods to complex survey data may produce misleading graphs, due to the effects of, *e.g.*, differential sampling rates or intracluster correlation. Since a given inverse sample is a simple random sample from the original population, the above mentioned problems would not arise when standard graphical methods were applied to data from a single inverse sample.

However, for inverse samples with small or moderate m , a scatterplot from a single inverse sample may not suffice for many purposes. An alternative approach would be to use several inverse samples in conjunction with local smoothing methods, *e.g.*, bivariate density estimation. For purposes of optimization, it may be useful to consider adjustment of some features of standard (simple random sample based) bivariate density estimators (*e.g.*, bandwidth) to account for unconditional correlation across the multiple inverse samples. Within this context, note that at a given point on the plane, a customary (simple random sample based) density estimator can be viewed as a solution to an estimating equation. Consequently, it would be of interest to study specific ways in which the RSB results on estimating equation methods may shed light on efficient approaches to bivariate density estimation based on inverse samples.

6. IDENTIFICATION RISK

As noted by RSB and HOS, a major potential attraction of inverse sampling is that it allows the computation of approximately design unbiased point estimators and variance estimators without explicit use of weights, stratum labels or cluster labels. This is of considerable practical interest in the preparation of public-use datasets because release of these types of design information can increase the risk that a sample unit can be identified by a data user. This in turn may constitute a violation of statistical agency pledges of respondent confidentiality. See, for example, de Waal and Willenborg (1997) and Chen and Keller-McNulty (1998) for detailed discussion of confidentiality issues associated with the release of weights.

In addition, in many household surveys in North America, strata and primary sample units are defined largely through geographical factors. For example, a primary sample unit in the U.S. is often a county or a group of contiguous counties. Release of nominally uninformative primary sample unit labels, accompanied by demographic and household-level observations Y , can lead to identification of the primary sample unit if the PSU-level aggregates of the observations Y vary in distinctive patterns that are publicly known. For example, a given county may have an unusual demographic profile, or may have a distinctive pattern of expenditures, *e.g.*, for natural gas or electricity.

For this reason, it would be of interest to evaluate the extent to which public release of multiple inverse samples may provide information that would allow a data user to reconstruct weights or PSU-level groupings that are informative. For instance, in keeping with comments by Mantel (2002), suppose that a given measured variable Y is reported on a continuous scale, and that for many responding units, the numerical value of Y is unique. Then (in keeping with the comments in section 2) matching of the reported Y values across a very large number g of multiple inverse samples would allow a data user to estimate the probability weights associated with a given respondent i . This in turn could lead back to the abovementioned identification problems considered by de Waal and Willenborg (1997) and Chen and Keller-McNulty (1998). For certain extreme cases, similar problems may arise with the

identifiability of primary sample units. The extent to which these issues are of practical concern depend on the relative empirical magnitudes of various error sources (including error induced by the use of finite g), and would be of interest to study for specific agency cases.

ACKNOWLEDGEMENTS

The author thanks Van Parsons, Fritz Scheuren and Al Zarate for many useful discussions of inverse sampling and its possible use in reduction of identification risk. The views expressed here are those of the author and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

ADDITIONAL REFERENCES

- CHEN, G., and KELLER-MCNULTY, S. (1998). Estimation of identification disclosure risk in microdata. *Journal of Official Statistics*. 14, 79-95.
- DUNCAN, G.T., and PEARSON, R.W. (1991). Enhancing access to microdata while protecting confidentiality (with discussion). *Statistical Science*. 6, 219-239.
- KORN, E.J., and GRAUBARD, B.I. (1995). Analysis of large health surveys: Accounting for the sampling design. *Journal of the Royal Statistical Society, Series A* 158, 263-295.
- MANTEL, H. (2002). Floor discussion at Statistics Canada Symposium, November 8, 2002.
- RUST, K., and KALTON, G. (1987). Strategies for collapsing strata for variance estimation. *Journal of Official Statistics*. 3, 69-81.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model-Assisted Survey Sampling*. New York: Springer-Verlag.
- SCHEUREN, F.J. (1997). Personal communication.
- DE WAAL, A.G., and WILLENBORG, L.C.R.J. (1997). Statistical disclosure control and sampling weights. *Journal of Official Statistics*. 13, 417-434.
- ZASLAVSKY, A.M., SCHENKER, N. and BELIN, T.R. (2001). Downweighting influential clusters in surveys: Application to the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*. 96, 858-869.

Comment

SUSAN HINKINS¹

Rao, Scott, and Benhin (RSB) have done an excellent job of summarizing our results on inverting complex samples and they have moved the subject substantially further with an impressive body of theoretical results. Their paper develops valuable new insight for statisticians who wish to consider, at the design stage, the option of using resampling techniques in the analysis. In this way, invertible designs can be used. As the authors point out, there are still many interesting problems to be considered.

We discuss here some specific points from the paper and also some of the open problems raised in our applied research into the employment of inverse sampling.

How to Use the Resulting Samples – The estimation of totals or means has the advantage that the combined and separate estimates are identical. Once one moves beyond “simple” estimation problems, there are many open questions as to the best use of the resulting samples, but combining the samples is a most reasonable approach. For a parameter such as a ratio, that is a function of totals, it would seem intuitive to calculate the best estimate of each total, and apply the function to the estimates and this is what we would recommend. In fact, because the ratio estimator is used in many situations, we did comment briefly on this in the 1997 HOS paper. However, RSB have made this point explicitly and in addition they have provided a coherent methodology for the estimation of variance from combined samples. This provides researchers with valuable tools for applying the inverse sampling techniques to a wider range of problems.

In Hinkins, Oh and Scheuren (1995), we considered the use of inverse sampling for the problem of calculating tests of independence from a 2x2 contingency table when the data come from a stratified sample. Contingency table analysis and regression analysis were both developed largely in the IID world and, therefore, adjustments are needed to use them in complex survey settings. We drew multiple simple random samples and calculated the simple Pearson chi-square test from the combined data. As the number of samples increases, the probability of rejecting the null hypothesis also increases, so one cannot take an arbitrarily large number of simple random samples. The problem was how to calibrate the tests, so that the desired level (e.g., a 0.05 significance level for example) is achieved. Preliminary results indicated that one could determine the number of simple random samples to combine to achieve the desired level for the test, and using the Pearson chi-square on the combined samples compared well to the Fellegi (1980) methodology applied to the original stratified sample, while perhaps being more user friendly.

In work that Hinkins, Liu, and Scheuren presented at the 1998 Statistical Society of Canada Conference, simulation results were shown for regression fits to inverse samples from a complex design (stratified median balanced design). In this case, the original design selected 100 replicates; in each replicate, one observation was selected from each of six strata, so that the observations were median balanced (Liu 1999). The selection was with replacement across replicates. The inverse sample consisted of selecting one unit from each replicate. We looked at regression fits to individual inverse samples, and at the regression fit to the combination of several inverse samples. The population regression line had a slope of 0.842 and $R^2 = 0.64$. Using single inverse samples, the estimated slopes ranged from 0.70 to 1.13. Combining six inverse samples, the estimated slope was 0.845 with $R^2 = 0.64$.

Variance Estimation – The estimation of variance in the HOS 1997 formulation is an interesting problem because the samples are not unconditionally independent. In our 1997 paper we suggested for ratio estimates that if the combined sample is sufficiently large so that a Taylor Series approximation is acceptable, then the “usual” approximation to the variance for a ratio could be used. That is, the variance could be estimated using the approximation

$$\text{Var}(\hat{R}) \doteq \frac{1}{\bar{X}^2} \text{Var}(\bar{e}) \text{ where } R = \frac{Y}{X} \text{ and } e_i = y_i - R x_i$$

The estimated variance for the ratio estimate based on the combined samples can then be calculated in the “usual” manner as

$$\text{var}(\hat{R}_c) = \text{var}\left(\frac{\bar{y}_c}{\bar{x}_c}\right) = \frac{1}{\bar{x}_c^2} \text{Var}(\bar{e}_c)$$

where $\bar{e}_c = (1/g) \sum_{j=1}^g \bar{e}_j$ and the mean in the j^{th} resample is $\bar{e}_j = \bar{x}_j - \hat{R}_c \bar{y}_j = \bar{x}_j - (\bar{y}_c / \bar{x}_c) \bar{y}_j$.

Using the estimate of variance generalized by the RSB equation (3.4) to estimate the variance of \bar{e}_c results in the following variance estimate for the combined ratio estimate:

$$\text{var}(\hat{R}_c) = \frac{1}{\bar{x}_c^2} \left(\frac{1}{g} \sum_j \left(\frac{1}{m} - \frac{1}{N} \right) s_{je}^2 - \frac{1}{g} \sum_j \bar{e}_j^2 \right)$$

$$\text{where } s_{je}^2 = \frac{1}{m-1} \sum_{i=1}^m (e_{ji} - \bar{e}_j)^2$$

¹ Susan Hinkins, Senior Statistician, National Opinion Research Center, 1122 South Fifth Ave, Bozeman, MT 59715. E-mail: hinkins-susan@norc.net.

As one would expect, this is the same variance estimator for the combined ratio estimate as Rao, Scott and Benhin construct using their estimating equations technique.

The use of the combined resampled samples for estimating regression coefficients is also addressed by RSB. They have developed a variance estimate, using the estimating equations technique, which appears to work well and further expands the possibilities for the use of resampling techniques. Their result also allows further research on the properties of the estimated variance in combined samples.

In the RSB regression example, it is not clear whether the variance estimate for B_0 has converged. The question of convergence of the estimates of error for nonlinear parameters is an interesting one, especially since these estimates are likely to be used to calibrate the process. (By calibration we mean the determination of when "enough" samples have been drawn, based on the desired use.) In the case of estimating a parameter, the only information available for calibration may be the comparison of the combined estimate, for example, to the original estimate from the complex design, and the comparison of their estimated standard errors. That is, while we may know that the variance will converge, only the estimates of variance are available for calibration.

Consider the following example where the inverse sample algorithm is used to invert a design with three strata and the minimum stratum sample size is two. Therefore, each re-sample is of size $m = 2$ and one would not expect fast convergence. Two ratios are estimated. Using 1,000 re-samples, the point estimates from the combined samples are within $\pm 1.0\%$ of the original estimates; using 10,000 re-samples the point estimates are within $\pm 0.3\%$ of the original estimates.

The estimates of the standard errors behave quite differently, however. For each parameter, Table 1 shows the ratio of the estimated standard error for the combined simple random samples to the estimated standard error of the original stratified estimate. The estimate of variance for the combined estimates was calculated using the method described above.

Table 1
Ratio of Estimated Standard Errors: Combined Estimate to Original Estimate

	Parameter	1,000 samples	10,000 samples
Totals	X_1	1.22	1.03
	Y_1	1.21	0.99
Ratio Estimate	$R_1 = Y_1/X_1$	1.07	1.07
Totals	X_2	1.02	0.95
	Y_2	0.94	0.93
Ratio Estimate	$R_2 = Y_2/X_2$	0.46	0.98

Using 1,000 re-sampled simple random samples, the estimated standard error of the combined estimate of X_1 is 22% larger than the estimated standard error for the original stratified estimate of X_1 . Incidentally, this was not surprising to us. With 10,000 re-samples, the standard error for the combined estimate is reasonably close to that of the original stratified estimate. Similar results are seen for the estimate of Y_1 . The standard error for the combined estimate of the ratio R_1 however converges more quickly, and appears to be relatively stable.

Consider the second set of variables. This time the standard errors for the combined estimates of the totals X_2 and Y_2 appear to have converged with only 1,000 samples. On the other hand, the standard error for the estimate of R_2 is severely under-estimated, as compared to the standard error of the original stratified estimate. An additional 9,000 draws, however, increases the estimated standard error for the ratio so that it is approximately equal to that of the original estimator.

Clearly, more analysis on the use of inverse sampling and the variance estimation for ratio and regression estimates is needed. Also, this example points out that the calibration of the inverse sample must consider all parameters of interest.

The remainder of the discussion considers two areas of interest where inverse sampling may be useful: providing public use data, and modeling or regression analysis. These two problems also illustrate two general types of data usage that may require different approaches to calibration.

Public-Use Data – The goal of using inverse sampling may be to provide public use data that will give substantially similar estimates as the estimates from the complex design, while permitting implementation of commonly available data analysis procedures using traditional computer software. If inference based on inverse-sample techniques can be demonstrated to be consistent with full complex-sample techniques, then data users with limited computer resources can perform select design-based analyses using mainstream statistical software. The results in the RSB paper expand the theory, providing conditions where the use of such resampling techniques is applicable.

A necessary feature in public-use data is the protection of confidentiality. For federal statistical agencies in the United States, public use files have been one of the responses to achieving the goal of "openness" (e.g., Duncan, Jabine and deWolf 1993). However, the growing electronic availability of data of all sorts through the Internet and the advances in record linkage software can be seen to endanger this openness (e.g., Doyle, Lane, Theeuwes and Zayatz 2001).

The goals of public use data can come into conflict when, for example, the information on the nature of the sample selection must be provided, implicitly or explicitly, for the calculation of design-based variances, but this information significantly increases the likelihood of

identifying an individual. In many surveys, geographical location plays an important part in the sampling, but the finer details of the geographical sampling structures cannot be released with the data without endangering the confidentiality of the individuals. If the geographical sampling structures are deleted to maintain confidentiality, then the data become difficult to analyze using the standard design-based methods. In this case, the use of inverse sampling would allow the release of data without the finer details of the geographical structures, for example, while still allowing analysis using the standard methods.

For example, the US National Health Interview Survey (NHIS) uses state-level stratification and selects counties and metropolitan areas for the sample. A public use file is released for the NHIS data in a form where the complex sample structure is simplified to that of a stratified design with two PSUs imbedded within each stratum. The original design strata and PSUs were masked in part using some of the techniques discussed in Eltinge (1999) and Parsons and Eltinge (1999). This masked "2 PSUs per stratum" design can be used to calculate variances. We investigated the NHIS design to see if inverse sampling was applicable for providing public use data (Hinkins and Scheuren 2001) and we found that it was not possible to invert the design down to the level of detail that was useful to data analysts. We still believe that inverse sampling can be an attractive option for providing public use data sets, when the design is invertible. It is not necessarily a viable option, however, unless its use is anticipated in the original design, so that invertible designs are used.

Another possible use of inverse sampling should be mentioned with respect to this example. For analytical domains covering most of the strata, the variance estimators from the NHIS public use data will be stable, *i.e.*, the estimators have large associated degrees of freedom. But for subpopulations that are less geographically dispersed, that cover few strata, the resulting degrees of freedom may be very small, and the variance estimate may be quite unstable. In such instances, it may be possible to produce a more stable variance estimator by drawing many, many samples from the public use design. In this case, rather than providing public use data, the inverse sampling might be used as a "black box" variance calculator that would provide more stable variance estimators for rare items in the population.

Modeling and Graphical Applications – Inverse sampling can be used to provide data in a form that allows greater analysis potential. This may be particularly valuable when there are multiple uses for the data. A natural example grows out of our initial proposal for using a resampling approach for the Statistics of Income (SOI) stratified samples of corporate tax returns. The underlying population is highly skewed (a relatively few large units accounting for a large percentage of the total value) and in order to provide efficient estimates of annual totals, a highly stratified sample design is used. However for economists, another

important use of the data is modeling economic activity and developing tax models, which is not the same problem as calculating a finite population regression estimate.

Another such example, from EPA, is a large stratified sample of US lakes from which water chemistry measurements were made in order to provide background measurements relating to acid rain. These data were also of great interest to biologists who were interested in modeling certain aspects of the chemical and physical relationships.

Interpreting regression models in finite population sampling can be confusing. There are many well thought-out approaches to regression in a complex sample setting, but the simplified rule of thumb is that you generally can't ignore the design structure (for example the sample weights.) To analysts interested in modeling the underlying parametric structure, this can seem counter-intuitive. And if the design is ignored, one can get the wrong answer unless either there are no missing regressors or the design is not confounded with regressors (both unlikely in our experience for complex designs). A simple random sample satisfies the second requirement. In the case of the SOI sample, if economists were interested in modeling the structure of the small to medium corporations, for example, then fairly large simple random samples could be generated from the stratified design. And a combination of multiple draws might provide a reasonable data base.

Finally, the use of graphical techniques in modeling and regression analysis is very important for understanding how a variable depends on other predictor variables. Even in the simple problem with one or two predictors of a dependent variable, the graphical display of relationships using weighted sample data is difficult. The analysis of residuals and the detection of outliers are more difficult with weighted data. Graphing is a powerful tool for extracting information from data. This would seem to be an area where the use of inverse sampling specifically for producing simple random samples should be considered.

As RSB rightly note in their conclusions, there are still many opportunities for further research and analysis. Their paper makes significant steps in advancing the theory and the application potential for the use of resampling procedures, opening doors to more opportunities.

ADDITIONAL REFERENCES

- DOYLE, P., LANE, J., THEEUWES, J. and ZAYATZ, L. (2001). *Confidentiality, Disclosure and Data Access*. North-Holland: New York.
- DUNCAN, G., JABINE, T. and DEWOLF, V. (1993). *Private Lives and Public Policies*. National Academy Press: Washington.
- ELTINGE, J.L. (1999). Use of stratum mixing to reduce primary-unit-level identification risk in public-use survey datasets. *Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference*.

- FELLEGI, I. (1980). Approximate tests of independence and goodness of fit based on multistage samples. *Journal of the American Statistical Association*, 75, 261-268. See also Scheuren, F. (1972), Topics in Multivariate Finite Population Sampling and Data Analysis: George Washington University Doctoral Dissertation.
- HINKINS, S., and SCHEUREN, F. (2001). *Increasing Public Accessibility to National Health Interview Survey Data (NHIS) Using Inverse Sampling*. Report prepared for NCHS under a Professional Services Contract.
- HINKINS, S., OH, H.L. and SCHEUREN, F. (1995). Using an inverse sampling algorithm for tests of independence based on stratified samples. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- HINKINS, S., PARSONS, V. and SCHEUREN, F. (2000). Increasing Public Accessibility to Complex Survey Data by Using Inverse Sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- HINKINS, S., LIU, Y. and SCHEUREN, F. (1998). Presentation at the Annual Statistical Society of Canada Meeting in June 1998.
- LIU, Y. (1999). *Balanced Sampling Design: An Improvement over the Classical Sampling Design*. Ph.D Dissertation. The George Washington University.
- MULROW, J., and SCHEUREN, F. (1998). The Confidentiality Beasties. *Turning Administrative Systems into Information Systems*. Internal Revenue Service.
- PARSONS, V.L., and ELTINGE, J.L. (1999). Stratum partition, collapse and mixing in construction of balanced repeated replication variance estimators. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Response From the Authors

1. INTRODUCTION

We thank the discussants, John Eltinge and Susan Hinkins, for their insightful comments and for suggesting some topics for further research on inverse sampling. In our rejoinder, we will attempt to address some of the issues raised by the discussants.

Our research on inverse sampling was motivated by the pioneering work of Hinkins, Oh and Scheuren (1997) (henceforth HOS). The latter authors developed several inverse sampling algorithms and provided some applications. They also noted the potential of inverse sampling in providing public-use microdata files, consisting of multiple simple random subsamples, that can be used to make valid inferences, such as regression and categorical data analysis, and to develop graphical displays of the data. The main contributions of our article is to provide some theoretical support (Theorems 1 – 5) and to develop the combined estimating equations (CEE) approach (section 5) to handle a variety of analyses of the data, such as linear and logistic regression, even when the subsample sizes are small. We have developed a linearization inverse-sampling variance estimator (equations (5.17) and (5.20)) that can be computed from the microdata file, and provided conditions for its convergence to the full-sample linearization variance estimator as the number of subsamples, g tends to ∞ .

(i) Point estimation of a total

In the context of estimating a total $\theta = Y$, we proposed the inverse-sampling estimator \hat{Y}_g given by (4.1) and showed that as $g \rightarrow \infty$, it converges to the full-sample Horvitz-Thompson estimator under the condition $\hat{\pi}_i(s_0) = \hat{\pi}_i$ for all $s_0 = i$ (see Theorem 3). Eltinge raised the important issue of improving the efficiency of \hat{Y}_g , for a given g . To this end, he suggested that single inverse sampling may be viewed as special type of two-phase sampling, and that using this analogy one could implement ratio or regression-based inverse-sampling estimators by constructing a public-use data set consisting of g subsamples, $\{(y_i, \mathbf{x}_i); i \in s_j^*, j = 1, \dots, g\}$, supplemented by the full-sample estimated totals \hat{X} for a vector of auxiliary variables, \mathbf{x} . For example, a ratio inverse-sampling estimator is given by $\hat{Y}_{rg} = (\hat{Y}_g / \hat{X}_g) \hat{X}$, where \hat{X}_g is the inverse-sampling estimator of the total X . Eltinge remarked that some additional auxiliary information may be required for variance estimation. It would be useful to pursue Eltinge's suggestions; one of us (E. Benhin) is looking into variance estimation. Benhin is also studying the analogues of full-sample calibration (or generalized regression) estimators constructed from multiple inverse samples (subsamples).

Eltinge also noted that in some cases the full-sample weights are adjusted to avoid problems with variance inflation induced by influential observations. He raised the question whether it is possible to modify the inverse sampling algorithms such that the resulting inverse-sampling estimator, say \hat{Y}_g , converges to the adjusted-weight full-sample estimator, say, \hat{Y} , as $g \rightarrow \infty$. This appears to be a challenging problem, but it may be possible to achieve approximate solutions.

(ii) Nonlinear parameters

In section 3 we considered a “separate” inverse sampling estimator, $\hat{\theta}_g$, of a nonlinear parameter θ , such as a ratio of totals $\theta = Y/X = R$, and noted that $\hat{\theta}_g$ can lead to large bias if the subsample size, m , is small. This is due to the fact that the bias of $\hat{\theta}_g$ is of the order m^{-1} . In her discussion, Hinkins noted that HOS were in fact aware of this problem and that HOS commented briefly on estimating the ratio R (page 18 of HOS). In particular, HOS suggested the estimation of the numerator Y and the denominator X separately, leading to the “combined” inverse-sampling estimator, $\hat{R}_{gc} = \hat{Y}_g / \hat{X}_g$, which follows as a special case of our CEE approach (see section 5.2). In section 5.1, we have spelled out the combined approach of HOS for the ratio R , at the suggestion of the Associate Editor, Fritz Scheuren.

(iii) Approximate variance estimator

Eltinge noted that approximate full-sample variance estimators, such as those based on stratum collapse, have been proposed in the literature and that it may be possible to develop inverse sampling procedures based on the approximate “variance estimation design” rather than the original sampling design. Such procedures may lead to larger subsample sizes, m . For example, in the case of stratified two-stage sampling with two clusters per stratum, we have $m = 2$ and m can be increased by stratum collapsing. This in turn may require a smaller number of subsamples, g , compared to the number of subsamples for the original design. Alternatively, for a given g , we may be able to obtain a more stable variance estimator, provided the full-sample approximate variance estimator is deemed to be satisfactory.

For PPS sampling without replacement, practitioners often assume that the sampling was with replacement to estimate the variance. In this case, HOS noted that “an inverse algorithm would exist to the same order of approximation as was being assumed to estimate variances” (page 16 of HOS).

(iv) Number of subsamples

The stability of the inverse-sampling variance estimator depends on the number of subsample, g , drawn from the full-sample and the function (or parameter) being estimated. For smaller g , the variance estimator can even take negative values. Also, when m is very small (as in the case of stratified two-stage sampling with two clusters per stratum), we will need a very large g to obtain a stable inverse-sampling variance estimator. We can increase m either by the approximate methods noted in (iii) or by drawing stratified random subsamples, provided confidentiality requirements or other considerations do not preclude the use of stratified subsamples.

Hinkins noted that the number of subsamples, g , may be determined by "calibrating" the inverse-sampling estimates and variance estimates to the corresponding full-sample values, and that the resulting g might vary significantly across parameters of interest. To illustrate the latter point, Hinkins studied the case of three strata and minimum stratum sample size of two, and computed the ratio, r , of the inverse-sampling variance estimator to the full-sample variance estimator for two ratios $R_1 = Y_1/X_1$ and $R_2 = Y_2/X_2$. Hinkins showed that the use of $g = 1,000$ subsamples leads to poor calibration for R_2 ($r = 0.46$ compare to $r = 0.98$ with $g = 10,000$). This result is somewhat surprising, but it could be attributed to the instability of the inverse-sampling variance estimator with subsample $m = 2$. Hinkins noted that the inverse-sampling variance estimator for the intercept term B_0 in our Table 2 (denote CEE) may be behaving somewhat erratically as g increases. We agree with her, but it is difficult to address the question of convergence for nonlinear parameters such as B_0 . Clearly, we need more work on the choice of g for variance estimation under inverse sampling. Fritz Scheuren noted in private correspondence that "the data user does know, however, what the main users are going to do, so g can be chosen with the important parameters in mind. But, of course, not all".

(v) Analysis of survey data

Computations of valid standard errors of parameter estimators from a full-sample microdata set may not be feasible in the context of stratified multistage sampling without the identification of clusters and strata on the data file. Even when the necessary information for standard error calculations is available on the data set, an analyst may lack appropriate complete-survey software for many (or all) of the intended analyses, as noted by Eltinge. On the other hand, valid standard errors may be obtained via the CEE approach using microdata files containing multiple simple random subsamples without the need for survey weights, clusters identifiers, *etc.* Moreover, as noted by Eltinge, the additional computational steps for implementing the CEE approach "impose a relatively low incremental burden on the analyst, or can be absorbed into the analytic software that is transparent to the analyst". However, we need further

work on providing the necessary enhancements to standard software in order to implement the CEE method in practice.

Hinkins, Oh and Scheuren (1995) combined the subsamples to test independence in a 2×2 contingency table. Their Pearson chi-squared statistic is of the form

$$X^2 = (gm) \sum_{i=1}^2 \sum_{j=1}^2 (\hat{P}_{ijg} - \hat{P}_{i\cdot g} \hat{P}_{\cdot jg})^2 / (\hat{P}_{i\cdot g} \hat{P}_{\cdot jg}),$$

where \hat{P}_{ijg} is the inverse-sampling combined estimator of the (i, j) -th cell proportion P_{ij} calculated from g subsamples each of size m , and $\hat{P}_{i\cdot g} = \sum_j \hat{P}_{ijg}$, $\hat{P}_{\cdot jg} = \sum_i \hat{P}_{ijg}$. It is clear from the form of X^2 that it increases with g so that the probability of rejecting the null hypothesis also increases with g . Hinkins, Oh and Scheuren (1995) noted that it may be possible to determine the number of subsampling, g , to combine to achieve the desired test level using X^2 . This idea looks interesting, but actual implementation of the method needs further study, especially for testing hypotheses in multi-way tables. Instead of using this approach, it is possible to develop first- and second-order Rao-Scott corrections to X^2 by using the multiple subsamples to implement Rao and Scott (1984) corrections, based on the concept of design effects. These adjusted X^2 will be valid for any g . Benhin is currently studying the Rao-Scott corrections in the context of inverse sampling. As $g \rightarrow \infty$, the corrected X^2 will converge to the Rao-Scott adjusted X^2 based on the full-sample.

(vi) Graphical displays and modeling

Direct application of standard methods for statistical graphics and modeling to complex survey data may produce misleading graphs and models, as noted by Eltinge, due to the effects of clustering, unequal weights, stratification and other features of the survey data. On the other hand, it is appropriate to apply standard methods to data from a single inverse sample (or subsample), provided the subsample is simple random sample unconditionally. However, the subsample size, m , is typically small and hence the subsample data set is not informative for graphical displays or modeling. The size of the data set may be increased to gm by combining the g subsamples, but the application of standard methods (*e.g.*, scatter plots) to the combined data set can produce misleading displays and inferences because the subsamples are unconditionally correlated. Eltinge made some useful suggestions on accounting for the unconditional correlation in the context of bivariate density estimation, but much work remains to be done in the area of statistical graphics and modeling using multiple inverse samples.

(vii) Confidentiality of microdata

As noted by Eltinge, a major potential attraction of inverse sampling is that it allows the calculation of point estimators, standard errors, *etc.* from the microdata file, consisting of multiple subsamples, without the knowledge of weights, cluster labels or stratum labels. This feature

allows the reduction of identification risk induced by the knowledge of cluster labels etc. It could be a challenging task to evaluate the extent to which the data file of multiple subsamples allows data users to reconstruct weights or cluster labels. Note that the characteristic values reported on the data file of inverse samples are real in the sense of corresponding to the values in the full sample.

If the full sample is a PPS cluster sample and the subsamples are obtained by selecting one element from each cluster, then cluster identification may be avoided by first randomly permuting the data vectors within each subsample and then reporting the permuted subsamples. The CEE approach is invariant to permutations of data vectors within each subsample.

It should be noted that the confidentiality protection provided by the data set with multiple subsamples is never more than the protection provided by a simple random full sample. Various methods have been proposed in the literature for limiting disclosure in microdata obtained from simple random sampling, such as microdata masking (see e.g., Cox 1994). We can use similar methods on the data set with multiple subsamples, if necessary. Raghunathan, Reiter and Rubin (2002) proposed multiple imputation for statistical disclosure limitation in the context of simple random sampling. The basic idea behind their proposal is to simulate multiple copies of the population by imputing for the nonsampled values using an imputation model based on auxiliary variables available for all the units in the population and then releasing a random sample from each of the synthetic populations. They used a parametric model-based approach and an approximate Bayesian bootstrap method for imputing the nonsampled values. The parametric approach protects confidentiality more effectively since the imputed values do not contain observed records, unlike the approximate Bayesian bootstrap, but it is far more susceptible to misspecifications of the imputation models. Note that the Raghunathan *et al.* (2002) method is fundamentally different from our method for complex full-samples. However, it is interesting to note that the variance estimator of Raghunathan *et al.* is given by

the variance between the imputed data estimators **minus** the average of the imputed data variance estimators, whereas our variance estimator (3.3) is given by the average of the subsample variance estimators **minus** the variance between the subsample estimators. In the case of multiple imputation for missing data, the variance estimator is given by the average of the imputed data variance estimators **plus** the variance between the imputed data estimators, treating the imputed values as the true values.

(viii) Concluding remarks

As noted by Hinkins, inverse sampling is not necessarily a viable operation unless its use is anticipated at the full-sample design stage to permit the use of invertible designs. Currently, we do not have inverse sampling procedures for several commonly used full-sample designs. For example, consider single stage cluster sampling with probability proportional to a measure of cluster size M_i , not necessarily equal to the actual cluster size M_i . In this case, we cannot apply the algorithm in Case 3 of section 2 to get a simple random subsample.

Further work is clearly needed on developing suitable algorithms to achieve exact matching or at least approximate matching with simple random sampling or stratified random sampling. As Fritz Scheuren noted in private communication, "this stuff is fun, but lots of fence to paint yet".

We thank the Associate Editor, Fritz Scheuren, for his interesting observations on our rejoinder.

ADDITIONAL REFERENCES

- COX, L.H. (1994). Matrix matching methods for disclosure limitation in microdata. *Survey Methodology*. 20, 165-169.
- RAGHUNATHAN, T.E., REITER, J.P. and RUBIN, D.B. (2002). Multiple imputation for statistical disclosure limitation. Technical report, Department of Biostatistics, University of Michigan, Ann Arbor.

The Accuracy and Coverage Evaluation: Theory and Design

HOWARD HOGAN¹

ABSTRACT

This paper discusses both the general question of designing a post-enumeration survey, and how these general questions were addressed in the U.S. Census Bureau's coverage measurement planned as part of Census 2000. It relates the basic concepts of the Dual System Estimator to questions of the definition and measurement of correct enumerations, the measurement of census omissions, operational independence, reporting of residence, and the role of after-matching reinterview. It discusses estimation issues such as the treatment of movers, missing data, and synthetic estimation of local corrected population size. It also discusses where the design failed in Census 2000.

KEY WORDS: Dual system estimation; Census adjustment; Undercount.

1. INTRODUCTION

The U.S. Census Bureau attempted to correct the initial Census 2000 population figures for measured net undercount (U.S. Census Bureau 2000.) This correction was to be based on the Accuracy and Coverage Evaluation (A.C.E.). The A.C.E. is a post-enumeration survey based on the dual system estimator (DSE). Although seemingly well designed and well executed, the initial A.C.E. production estimates were badly flawed. The A.C.E. produced an estimate of 3.3 million net undercount (378,000 s.e.). This contrasts sharply with the current demographic analysis estimate of only 340 thousand (Robinson 2001) as well as a later revised survey estimate of a 1.3 million overcount (542,000 s.e.) (U.S. Census Bureau 2003).

This paper discusses both the general question of designing a post-enumeration survey (PES), and how these general questions were addressed in the U.S. Census Bureau's plans for the A.C.E. Where applicable, it discusses where the assumptions underlying the design failed in 2000. Throughout, I will use the terms DSE and PES when a general question is discussed and A.C.E. for specific details of the U.S. 2000 design. The next section defines the dual system model as applied to census coverage measurement. Section 3 discusses the definition and measurement of census correct and erroneous enumerations. Section 4 presents the issues in defining and measuring omissions. Section 5 deals with small area estimation. The paper ends with a discussion of some of the problems encountered in implementing the A.C.E. together with some concluding remarks.

2. THE DUAL SYSTEM ESTIMATION MODEL

The use of the dual system model is well known either for measuring the completeness of vital events registration (Sekar and Deming 1949; Marks, Seltzer and Krotki 1974)

or for use in measuring coverage errors in census data (Marks 1979; Wolter 1986; U.S. Bureau of the Census 1985.) Application of the dual system model in the context of the 1990 Census, including the issue of census adjustment, is documented in Hogan (1992, 1993.)

The standard Petersen (1896), Sekar-Deming or dual system estimator (DSE) can be expressed as:

$$\hat{N}_{++} = N_{+1} (N_{1+} / N_{11}) \quad (1)$$

where

N_{11} is the number of people counted in both the census and the survey,

N_{+1} is the number of people correctly counted in the census,

N_{1+} is the number of people counted in the survey, and N_{++} is the total number of people.

That is, the total population is estimated by the number captured in the census multiplied by the ratio of those in the survey to those in both systems (*i.e.*, the inverse of the coverage rate of the census, as measured by the survey).

The DSE will yield a direct estimate of the population of class j , as well as any sum of classes. The class j might be the household population of a state, of a district, of an ethnic group, or perhaps of an ethnic group within a state.

Requirements for estimating small or local populations, for example, age by sex, by race, by town, often far exceed the capacity of even a very large sample. To meet this need, the DSE is combined with a synthetic assumption to produce estimates for areas of geography smaller than that defined by the domain j . The synthetic estimator assumes that a proportion or ratio measured at an aggregate level applies equally to all sub-groupings (Gonzalez 1973; Gonzalez and Hoza 1988.) Using a synthetic assumption, we write

$$\hat{N}_{jkh}^s = CCF_j C_{jkh} \quad (2)$$

¹ Howard Hogan, Chief, Economic Statistical Methods and Programming Division, Census Bureau, Washington, D.C. 20233.

$$CCF_j = \frac{\hat{N}_j}{C_j} \quad (3)$$

where,

\hat{N}_{jkh}^s is the estimated population in domain j , available at the level of geography k and demographic subclass h .

CCF_j is the net coverage correction factor

\hat{N}_j is the DSE for domain j

C_{jkh} is the measure (usually census count) of the population in domain j available at the level of geography k and demographic subclass h , and

$$C_j = \sum_k \sum_h C_{jkh}. \quad (4)$$

C_j need not equal the number of people correctly included in the census (N_{+1}). N_{+1} is estimated from sample data and is not available for all small areas. C is normally the census count, including imputations and erroneous inclusions (duplicates, *etc.*).

Summing over group j and subclass h yields a measured population for the given geographic area k , we have

$$\hat{N}_k^s = \sum_j \sum_h CCF_j C_{jkh}. \quad (5)$$

For example, j may define all 0-17 year-old Asians in owner-occupied housing units while k may define Orange County, California, and h may define 11-year-old girls.

While this produces a small-area and small-group estimate, this calculation can generate fractions. The typical user of census data prefers whole person records. The U.S. Census uses controlled rounding and person record imputation to create integer number of person records for ease of tabulation and data acceptance.

3. MEASURING CORRECT ENUMERATIONS

3.1 Defining and Measuring Correct and Erroneous Enumerations

The first step in operationalizing Equation 1 is to define and estimate the set of individuals “correctly” in the census. In this context “correctly” has four dimensions:

1. Appropriateness
2. Uniqueness
3. Completeness
4. Geographic correctness

“Appropriateness” means that the person should be included in the census. People who die before or who were born after the census reference date (April 1 in the U.S.) are not part of the population (universe) to be measured. Similarly, records that refer to fictitious “people,” tourists, or animals are out-of-scope.

“Uniqueness” refers to the fact that we wish to measure the number of people included in the census, not the number of census records. If more than one record refers to a single person, the count of records must be reduced for purposes of the DSE.

“Completeness” means that the census record must be sufficient to identify a single person. If it lacks sufficient identifying information, we cannot determine whether the person was appropriately and uniquely included in the census, nor can we determine whether he or she was also included in the survey.

Although completeness is necessary for the DSE, the census count includes imputations and other incomplete enumerations. Census operations normally have a requirement for a “data-defined person record.” In Census 2000, the requirement was two characteristics where name counts as a characteristic. The name field must have at least three characters in the first and last name fields combined. The characteristics that are included in the counting are relationship to the householder, sex, race, Hispanic origin, and either age or year of birth. (Childers 2001)

When a record does not meet these requirements census processing substitutes (imputes) a data-defined record. Since the census processing identifies all these whole-person imputations, the quantities are known and need not be estimated. Traditionally, the number of whole person imputations is denoted by II , for “insufficient information.”

Additionally, there are person records that are acceptable for census processing but insufficient for use in the DSE. This group includes records with reasonably complete data but without a person’s name. Accurate matching or additional interviewing is not possible for these cases. For A.C.E. 2000, the definition for “sufficient information for matching” was complete name and two characteristics. (Childers 2001)

“Geographic correctness” means that people are included in the census where they should be included. Enumerations outside that defined search area (or areas) are counted in the census but not correctly included in the census. This area must be searched during the matching process as well as searched for census duplicates. As the number of addresses in the search area increases, the complexity of matching increases and the chance of matching error grows. This increased complexity and possible levels of error will affect both the matching between the survey and the census and the search for census duplicates. The more addresses that must be searched, the more likely a true match will be missed. Equally importantly, the chance of a false match increases. For example, the chance of finding two people with similar names and ages living in the same block is small. The chances of finding two such people in a large city is considerable.

Two dimensions must be defined to operationalize a search area: (1) correct location and (2) the search area around the correct location.

The “correct location” defines where, under the DSE residence rules, the person should be included in the census. These rules may differ from the rules used in the census. The only requirement is that the location be precisely defined and consistently applied during PES processing. More than one location may be defined as correct so long as the rule is consistently applied. However, usually only one location is defined as correct. This was the rule in the A.C.E.

In the 1990 PES and 2000 A.C.E. the Census Bureau adopted the following rule:

The person is correctly included in the census if he or she is included at the location where the person considers, at the time of the survey interview, to have been his or her usual residence as of April 1.

This definition generally follows the census rules. However, it makes an explicit allowance for the fact that the concept of “usual residence” is somewhat subjective. Because of this subjectivity, where the person considers his/her usual (April 1) residence may have changed by the time of the survey interview. This, by itself, does not bias the DSE. However, it does require consistent reporting of the “correct location.”

The second dimension of geographic correctness is the area of search around the correct location, *i.e.*, the search area. The concept of a search area is to accommodate errors in either the census or survey assignment of residents to a particular geography. It has the effect of lowering the variance and can, in some circumstances, lower the bias as well.

The A.C.E. used the following definition:

A person was correctly enumerated if the person was counted in the block cluster containing his/her usual residence; or if he/she was included by the census in the housing unit where he/she usually resides, and the housing unit was included in a block adjacent to the correct block cluster.

An important part of this design is that enumerations of people in the “wrong” location are to be classified as erroneous, whether or not the people are also enumerated in the correct location. Thus a person counted only once, but in the wrong location, should be measured, on average, as contributing one erroneous enumeration (in the wrong location) while being missed (one omission) in the correct location. This approach obviates the need to search widely for possible duplicates, but does require that the field interview determine a unique correct location for each person.

The definition of “correctly included” does not depend on the correctness of classification *j*. For example, if a person was really 19 years-old, but was counted in the census as 17, he/she is still considered as correctly included. This is discussed in section 5.2.

To estimate the number of people correctly included in the census, one must take a sample of all data-defined census enumerations. This sample is called the enumeration (or *E*) sample. Census whole-person imputations (II's) are not part of the *E*-sample frame.

To maximize correlation with the population sample (see below), the A.C.E. first defines a set of sample areas. These are either a single block or a group of contiguous blocks and are known as block clusters. If a block is sampled, all census records coded to that block, even incorrectly, fall into sample. If the block contains many census housing unit records it may be subsampled.

The records in the *E*-sample will be checked for completeness. Only records that meet the minimum completeness requirement can be considered as correctly enumerated in the census. Records are then searched throughout the search area to see if the person was counted more than once within the sample block (uniqueness). Duplicate search is done using computer-assisted clerical matching. If more than one record is found, the extra records are coded as duplicates.

Appropriateness and geographic location cannot be determined from the census enumeration alone, but require additional interviewing. If interviewing locates a member of the household, or an acceptable respondent who can confirm the person's existence and that the person had his/her usual residence there on April 1, the enumeration is accepted as correct.

If the respondent reports that the person did not live in the block or search area on April 1, the enumeration is excluded from the correct enumerations. This can occur when the person responded to the census but moved before April 1; the person moved in after April 1 but was enumerated by the census nonresponse follow up operation; or when a parent incorrectly reports a college student as living at home.

The interviewers may determine that the person never existed or was never associated with the block. These records are considered erroneous. It is difficult in some cases to prove that a “person” was not real, especially in a large block. The A.C.E. required the interviewers to find at least three knowledgeable respondents before coding a record as fictitious. However, since the person might have lived somewhere else in the block, it can be difficult in some situations to code the record fictitious.

An important source of error arises from the need to accept proxy responses to verify many enumerations. If the proxy reports a different “correct” residence than the person himself would, an enumeration could be miscoded, since the requirement of a unique “correct” residence would be violated. The A.C.E. used proxy interviews for households that moved between the time of the census and the time of the A.C.E. interviews. Even within a household, different members may hold different views of a person's “correct” residence on Census Day. Proxy respondents, both household and non-household, were responsible for many

of the errors in reporting residence in the A.C.E. and thus, the underestimation of census error.

After missing-data estimation and sample weighting, we can estimate the number of people correctly counted in the census as

$$N_{+1} = (C - II) \frac{CE}{N_e} \quad (6)$$

Where

C = Census total records, including imputed, duplicate, fictitious, *etc.* (the Census count),

II = number of whole-person census imputations,

CE = weighted estimate of appropriate, unique, complete and correct enumerations,

N_e = weighted E -sample estimate of total, including duplicate, fictitious, *etc.*

Occasionally, due to processing errors or timing constraints there may be a group of census enumerations that are excluded from both the E -sample processing and from the searching and matching process. Thus, while these records may be processed in time to be included in the official census results, they arrived too late to be included in coverage measurement processing. These cases are sometimes known as "Late Census Adds" (LCA). These cases can be handled analogously to the treatment of census whole person imputations, that is replace $(C - II)$ in Equation 6 with $(C - II - LCA)$. Excluding the LCAs will not affect the DSE of the true population if the number of matches is reduced proportionally to the number of census correct enumerations. Said another way, the assumption is that the probability of a LCA being excluded from the A.C.E. processing must be statistically independent of its inclusion probability in the A.C.E. This is, of course, the traditional dual system independence assumption. (See Hogan 2001 for the supporting theory.) Although there were 2.3 million LCAs in Census 2000, analysis of the A.C.E. results by Raglin (2002) showed a trivial impact on the final DSE results.

In situations where the number of whole person imputations (II) was small, $(CE/N_e - 1)$ would be a measure of census gross overcoverage. That measure, however, is a function of the operational definitions of "correctly enumerated" adopted by the coverage measurement design. Definitions adopted to produce a good measure of net coverage, especially with respect to completeness and geographic correctness, may differ from those most appropriate for studying the quality of Census field operations. In any case, Census 2000 included 5.8 million whole-person imputations, of which 1.2 million were for housing units where the interviewer was unable to obtain even the number of residents (see Table 1 in Nash 2001, and page ii of Wetrogan and Cresce 2001.)

4. MEASURING THE PROPORTION OF PEOPLE CORRECTLY ENUMERATED

Having defined the set of correctly enumerated people, the next step in the DSE is to estimate the census coverage rate, N_{11}/N_{1+} .

Conceptually, estimating the rate entails (1) taking a sample of people, (2) determining whether they should be enumerated in the census, and (3) determining whether they were, indeed, correctly enumerated, using the same definitions as were used to measure N_{+1} . If an unbiased sample can be drawn of people who should have been enumerated and, if we can determine whether they actually were correctly enumerated (included in the census), then the DSE will produce asymptotically unbiased estimates. If each step can be approximately correct, the results will approach an unbiased estimate.

The first step in the process is, normally, to draw a random area sample. The A.C.E. uses the same set of block clusters for this purpose that it uses to define the E -sample.

Interviewers then canvass the block and prepare an independent list of people who should have been enumerated. This list constitutes the population or P -sample. The (weighted) sum of the people on this list, denoted \hat{N}_p , estimates N_{1+} . However, it is not the number which is of interest, but the ratio of N_{11} to N_{1+} , which we approximate by the ratio of correct matches, \hat{M} , to \hat{N}_p .

Operationally, the "correctly enumerated" census records are searched to see if the P -sample people were enumerated. The (weighted) number who were matched (\hat{M}) estimates N_{11} .

The DSE model will work if we can approximate:

1. Operational independence
2. Consistent reporting of residence
3. Accurate matching
4. Homogeneity within post-stratum

4.1 Operational Independence

Operational independence is the easiest assumption to approximate, but still requires vigilance. In Census 2000, the A.C.E. sample was drawn and the housing units listed before the delivery of the census questionnaires. Although personal contact was minimal, some people may react differently to the census because of their inclusion in survey listing. Early telephone interviews were allowed for independently listed housing units linked to a census address with a completed census questionnaire. This operation occurred while census nonresponse follow up was still being conducted in the area. Personal visit interviewing took place concurrently with some census "coverage improvement" interviewing. Clearly, some contamination could occur. Great care was taken to prevent the same field staff from working the same area in both Census and A.C.E. and to prevent the sharing of information. Still, some people may react differently to the survey because

they were enumerated, for example, by a very polite or very surly enumerator. Others may believe that they have a duty to provide the information once, but not twice.

Operational independence must also be preserved in office procedures. Definitions of "nonresponse" or "sufficient information" are sometimes applied differently to matched and non-matched *P*-sample records. The A.C.E. guarded against unnecessarily introducing operational dependence by forcing the processing system to first decide whether a case is acceptable for matching and only then attempt matching. The philosophy is "Do not attempt to find a match unless you would be satisfied that, if no match is found, the person was not enumerated!"

Before beginning the matching, *P*-sample records first are reviewed for:

- (1) Appropriateness
- (2) Uniqueness
- (3) Completeness
- (4) Geographic correctness

The A.C.E. contained no obviously fictitious records. One important safeguard is the use of Computer Assisted Personal Interviewing (CAPI). The CAPI instrument makes falsification difficult by "time stamping" the interview and recording every key stroke. We have instituted a quality assurance process to minimize other sloppy or dishonest A.C.E. interviewing. In addition, one important exception to the "no follow up" rule are cases where A.C.E. fabrication is possible, *e.g.*, cases where no one in the household matches, implying possible fabrication.

Out of scope records, *e.g.*, group quarters, are screened out. Occasionally, survey duplicates occur and these are eliminated (uniqueness). Finally, if the survey interview does not meet minimal standards, the case is converted to nonresponse and is later imputed.

4.2 Consistent Reporting of Residence

To measure the number of people correctly in both systems, we must determine whether or not a *P*-sample person was correctly enumerated in the census. This is done by searching the correct census records in the area where the person should have been enumerated.

The same definition of geographic correctness must apply both to whether an enumeration (in the *E*-sample) was correct and to whether the person (in the *P*-sample) was correctly enumerated. Failure to make these concepts agree is termed "balancing error."

Specifically, we must have the same definition of "correct" location and the same search area around the correct location. Errors can result in both erroneous non-matches and erroneous matches. Difficulty comes primarily from two sources. First, both the *P* and *E*-sample accept proxy responses. Thus, even though the person might have a clear and consistent understanding of his usual residence, the proxy respondent may not. Secondly, the way in which

the question is posed in each interview could lead to different responses even from the same person. This might result in false non-match/not correctly enumerated status. On the other hand, if the person was incorrectly included by the census, we could incorrectly count the person as "correctly enumerated." Both errors clearly occurred on a relatively large scale in the A.C.E. (See section 6.)

The other dimension of geographic correctness is, again, the extent of search. The same area must be used to define the correct residence for determining both whether an enumeration was correct and whether a person was correctly enumerated. This is achieved by consistently applying the same search area definitions as in section 3.

4.3 Accurate Matching

The purpose of matching is to determine whether a person interviewed in the *P*-sample was also enumerated in the census within the defined search area. Much of the matching is now done by a computerized matching system. The system produces matches, possible matches, and non-matched cases. Repeated tests have shown that cases matched by the computer are nearly certainly correctly linked (Belin 1993). Nearly all clerical matching is now computer-assisted and largely paperless. This new system makes searching easier, including duplicate search. It restricts the codes clerks can apply to only those appropriate for the situation. The almost paperless system eliminated lost and misfiled A.C.E. questionnaires.

The first-level clerks were backed up by a team of 46 technicians. Training for these technicians began in September 1999. They were supported by a team of seven permanent analysts, most of whom have been matching for many years. Each level of matching acts as quality assurance for the level before. In addition, each level could refer problem cases to the next higher level. All matching was done in one location by one staff. The 1980 and 1990 matching operations were done in three and seven sites, respectively.

The use of the A.C.E. procedures for movers also greatly simplified the matching. Information about those who had moved was gathered from current residents. Under the procedures used in 1980 and 1990, movers were interviewed at their residence at the time of the PES interview. It was necessary then to code the reported correct Census Day residence to the correct census geography before beginning matching. This procedure was difficult, especially in rural areas. Mover matching was never before automated. In A.C.E., all matching, including for movers, was done in the *E*-sample block cluster or an adjacent block, using the same computer and computer-assisted clerical matching system. The change in the treatment of movers is discussed below.

4.4 The Role of After-Matching Reinterview

Some cases are sent to the field to gather further information after the initial matching is complete. This

after-matching reinterview is often termed "follow up interview."

The follow up interview process, like all PES activities, must fit into the overall framework of the DSE. Specifically, it must account for:

1. Appropriate, unique and correct response
2. Independence between census and survey inclusion probabilities
3. Balancing *P* and *E*-sample concepts
4. Search area and unique location matching rules
5. Treatment of missing data.

Follow up is only useful if it provides more accurate or consistent responses. Simply obtaining a different response is not justification. Since follow up takes place further from the census reference date than the initial interview, it is more difficult to obtain accurate responses. This is equally true for *E*-sample follow up and *P*-sample follow up. To provide better responses, follow up must use better resources, for example: (1) better respondents (household vs. proxy), (2) a better trained, supervised or quality-controlled interviewer, or (3) better questions or interview procedures.

The census data collection period extends from mid-March through mid-summer. Because of the huge scale of the operation, little emphasis is placed on verifying that the people were residents of the household on April 1. Quality assurance reinterview to prevent fabrication is minimal. Because of better training and supervision, and more complete questioning, the A.C.E. follow up interviewing can, in general, obtain more accurate information on residence and location than that gathered during the census process itself. Thus all non-matched *E*-sample cases were sent to follow up.

Follow up can, however, compromise independence. If all cases were sent to follow up, independence would not necessarily be compromised. However, cases that are matched during initial matching are seldom sent to follow up. To do so would stress the resources available for follow up. Instead, only non-matches or "possibly matched" cases are usually selected for follow up. This can introduce operational dependence.

The biases that can be introduced by follow up can occur even if the follow up interview was successfully conducted, since follow up may selectively change the defined "correct location" for non-matches but not for matches. If the follow up operation results in a non-interview, further biases can be introduced depending upon the missing data models applied to these cases.

Choosing cases for follow up requires balancing the need for accurate and consistent information with the need for independence. The *P*-sample only followed up cases when better information was likely. Cases sent to follow up included:

1. Possible matches, since with the information at hand the interviewers can resolve the situation,
2. Initial non-household proxy interviews that result in non-matches. Since we have not spoken to a household member, we have reason to doubt the accuracy,
3. Non-matched cases where, for the same housing unit, the census reports one family and the A.C.E. reports another. In order to ensure consistent reporting of Census Day address between the *P*-sample and the *E*-sample, these cases are sent out together,
4. Partial-household non-matches.

Cases that match and some other non-matched cases were generally not sent to follow up. For example, the A.C.E. did not follow up whole-household nonmatched cases where the census missed the unit, reported it as vacant, or could not obtain an interview (last resort information only).

4.5 Homogeneity Within Post-stratum

The DSE requires that the capture probabilities be independent for all individuals within estimation domains called post-strata. This is approximated by making the post-strata as homogeneous as possible with respect to the census capture probabilities, and then striving for as uniform as possible inclusion probabilities for the survey.

Dividing the population into many relatively small post-strata can increase within strata homogeneity. However, small strata can have high sampling variance and ratio bias. Ratio bias follows from the fact that the DSE is inherently a ratio estimator. This bias tends to decrease as the size of the post-stratum increases. In addition, our treatment of movers adds an additional ratio (see below). For this reason, we designed post-strata with a minimum expected sample size of 100.

For the A.C.E. we post-stratified based on the following variables:

1. Race / Hispanic Origin (7)
2. Age / sex (7)
3. Tenure (2)
4. Metropolitan area size and type of enumeration area (4)
5. Return rates (2)
6. Region (4)

where the number in parenthesis refers to the number of categories. More details on the post-strata are found in Haines (2001).

Coverage differences between racial and ethnic groups is well documented. (See for example Robinson, Ahmed, Das Gupta and Woodrow 1993; Hogan 1993.) Social, cultural, linguistic and economic differences may lead different racial and ethnic groups to react differently to the census procedures.

Demographic analysis and previous coverage surveys have demonstrated that people are differentially missed in different age groups and that the pattern is different for males and females. Most important in this pattern is young adults (Robinson *et al.* 1993.)

The importance of tenure was first measured following the 1980 Census and then implemented in the 1990 post-stratification. Those who live in owner-occupied houses are less mobile. They may feel that they have more of a stake in their community and thus, are more influenced by the census outreach program.

Metropolitan area size obviously affects housing patterns and is correlated with the way the Census Bureau builds its address lists. The combined variable "metropolitan area size and type of enumeration area" isolates differences in housing unit coverage. It may, in addition, measure some aspects of social and economic isolation.

The census return rate measures public cooperation with the census, an important predictor of coverage. It also measures directly the proportion of the enumeration that must be done in the census nonresponse follow up. One difficulty in this variable is that not all areas of the country are in the mailback universe. A small proportion is done by direct interview, and obviously have no "return rate." We have chosen to group these areas with "high" mail response areas.

Census Region picks up, among other things, broad differences in settlement patterns and housing stock. "Brown stone walk ups" are more common in the Northeast. Mobile homes are more common in the South.

Obviously, the complete cross-classifications can lead to very small cells. The maximum set of post-strata the sample was designed to support was 448. In fact, after collapsing small cells, there were 416 post-strata.

4.6 Treatment of Movers

People who move between the census reference date and the time of the survey interview present a challenge for designing a DSE for census application. First, people who move are more likely to be missed by the census and by the survey. Secondly, if a person has a different "usual residence" at the time of the survey than he did at the time of the census, one must decide where to sample him.

In the 1990 PES, movers were sampled where they lived at the time of the survey interview. We then searched the census records at, and only at, their April 1 usual residence. This is known as procedure B (Marks 1979). This approach requires both coding the address to the correct Census Day geography and then matching. These activities are complex and time consuming.

The A.C.E. used a different procedure known as procedure C. The A.C.E. estimated the number of movers by the number of people who moved into the sample blocks between April 1 and the time of the A.C.E. interview (in-movers). If the population was closed to international migration, deaths, movement to group quarters, *etc.*, then

the number of people who moved in must equal the number who moved out (out-movers). They are the same people in the population, if not in the sample. It is normally easier to find people where they are, so the measured number of in-movers is normally a better estimate of the total number of movers than the measured number of out-movers.

The proportion of movers who are correctly enumerated is estimated by matching the out-movers to the census records for the sample block and extended search area, if appropriate. The estimated number of correctly enumerated movers is then $\hat{M}_t = (\hat{M}_o / \hat{N}_o) \hat{N}_i$ where \hat{M} denotes the weighted number of correct matches; \hat{N} denotes the weighted population number; and the subscripts denote total movers (*t*), out-movers (*o*) and in-movers (*i*).

If we denote those who do not move by the subscript *n*, the overall coverage rate becomes

$$\frac{N_{11}}{N_{+1}} = \frac{\hat{M}_n + \hat{M}_t}{\hat{N}_n + \hat{N}_i}$$

The effect of procedure C is to increase the effective capture probabilities in the survey for movers and thus increase homogeneity of inclusion in the survey with respect to mover status (*i.e.*, mover/nonmover) (Griffin 2000).

There will be nonresponse and incomplete response at various steps. The goal of the missing data process is to improve the estimate of the number of people correctly counted (from the *E*-sample) or the estimate of the coverage ratio (from the *P*-sample). In designing missing data procedures, we choose methods that support the underlying DSE assumptions. Starting with the 1990 PES, the U.S. has estimated the probability a nonresponse record was correct rather than assigned a "zero/one" classification. (Schenker 1988, Belin 1993) The methods used for the A.C.E. are described in Cantwell and Ikeda (in this volume).

5. SYNTHETIC ESTIMATION

5.1 The Synthetic and Dual System Model

To this point, we have been dealing with the actual DSE. However, as noted in section 2, we use a synthetic estimator to distribute the measured net undercount to local areas and small groups.

In the A.C.E. the carrying-down was based on the same post-stratification variables as the DSE itself. The synthetic estimation is based on the assumptions that (1) the DSE estimates the true population, and (2) within post-strata, the true population is distributed proportionally to the pre-adjustment census count.

Clearly, at some level the second assumption can be only true with respect to the expected census counts. That is, even if within post-strata all people had identical probabilities of being enumerated in the census, we would observe different outcomes across blocks. The underlying

DSE explicitly models the undercount as a stochastic process.

As areas get larger, two things happen. First, the stochastic effect, or the random "block effect" begins to average out. Secondly, the effect of the actual undercount from a collection of blocks becomes positively correlated with the post-stratum's coverage correction factor. That is, the larger the area, the more the area's undercount determines the net correction factor.

The stochastic effect would be trivial for all but the smallest areas if Wolter's (1986) autonomous independence assumption held in practice, that is, if each person was included or missed independently of all other people. In fact, it is well known whole families are often missed or duplicated. Indeed, the whole building (or sometimes even block) might be missed or duplicated by the census address listing procedure. The failure of the autonomous independence assumption does not cause a bias in the dual system model as long as the underlying probabilities are equal within post-strata. This failure can mean that observed coverage for a block is inconsistent with the estimated undercount adjustment. However, as attention is turned to larger areas, the stochastic effect diminishes and is replaced with the problem of true heterogeneity of the underlying capture probabilities (see Haines 2001 for synthetic estimation details.)

5.2 Misclassification Error

In the discussion so far, we have accepted the post-stratum classification, j , as fixed. In practice, some people will be classified in different post-strata in the census and in the survey. For example, a woman may be reported as age 28 in the census and 31 in the survey, placing her in different post-strata.

Such misreporting is normally not important for matching. Name, address, month and day of birth, relation and household composition are far more important than age, race or sometimes even sex. So, assuming a match, in the above example we would have one correctly enumerated 28 year-old in the E -sample and one correctly enumerated 31 year-old in the P -sample. Misclassification can be seen to have two effects. To the extent the true undercount probabilities are homogeneous with respect to the true characteristics, misclassification introduces heterogeneity (and heterogeneity bias) into the observed estimation cells. This is true even if reporting is consistent between the census and the survey, because it can introduce unobserved subgroups within post-strata where the probabilities of inclusion in each system are correlated.

Inconsistent reporting between the census and the survey poses a problem for the synthetic estimator as well as for the DSE. This is easily seen by ignoring census imputations and erroneous enumerations. In this case, the coverage correction factor is the inverse of the matching rate (N_{11j}/N_{1+j}) where j represents the post-stratum. If the classification into the post-strata is inconsistent between the

census and survey, we would be applying the rate, estimated from one group, to a somewhat different group. While misclassification may be ignorable at the poststratum level, it may be important locally. The A.C.E. protected itself against the general problem by avoiding, when possible, post-stratum definitions based on variables with high reporting variability.

6. FAILURE OF THE A.C.E. DESIGN AND CONCLUDING REMARKS

In spite of being seemingly well designed and well executed, the A.C.E. failed to even approximately measure the coverage error in the 2000 U.S. Census. The chief reason seems to have been a failure of the assumption of consistent reporting of Census Day residence. In other words, depending upon when and where and with whom the interview was conducted two or more residences were reported as the correct one for a large number of people in sample.

We know that this happened because, after the both the census and the A.C.E. were completed, we were able to search and match nationally. This allowed us to search for census duplicates, even when the pair was miles apart. This was possible because, for the first time, practically all names in the census were data captured. (See Fay 2002; Mule 2001, 2002.) We could see, for example, how many of the people who were classified by the A.C.E. E -sample as "correctly enumerated" were also enumerated somewhere else, including at an other household or in a group quarters.

In one study, of the 1.3 million (weighted) E -sample people linked to a duplicate census enumeration outside the search area, only 14 percent were coded as erroneous enumerations by the A.C.E. (Feldpausch 2001, Table 1.) Since the A.C.E. E -sample was a random sample, one would expect that for any pair of duplicates it would pick up the erroneous enumeration roughly half the time.

Another 521 thousand E -sample cases (weighted) were linked to census enumerations in group quarters. Of these, only 31 percent were classified as erroneous by the A.C.E. (Feldpausch 2001, Table 3.) Roughly half, 271 thousand, of these linked E -sample cases were linked to an enumeration in a college dormitory. Under census residence rules, those living in a dormitory should be counted there, and not at home. However, the A.C.E. classified only 45 percent of these E -sample cases as erroneous enumerations. Since the proportions coded as correctly enumerated by the A.C.E. are significantly different from what would be reasonable, one must conclude that the A.C.E. had a strong tendency to misclassify enumeration status. Interestingly, many of these misclassified cases, the exact number is hard to determine, must have been A.C.E. matches. This is certainly due to the tendency of respondents to confirm people as living at an address who should be counted as living somewhere else.

We now have clear evidence that large number of parents of college students living in dormitories will consistently report their child as living at home even though census instructions clearly say not to. Further, both parents in a "joint custody" situation may consistently report the child as living in each of two households. Neighbors, no doubt, will report someone as "living there" who is in fact away at college, in the military, in jail, or at a second home. This misreporting occurred in spite of the numerous, detailed and specific probing questions about usual residence asked by the A.C.E.

The extended search for census duplicates discussed above formed the principal evidence for A.C.E. error. However, other evidence was also gathered, including a re-interview study. These evaluations are discussed in detail in the Census Bureau's "Executive Steering Committee on A.C.E. Policy" (ESCAP) documentation. (See ESCAP I 2001, ESCAP II 2001).

The results of these evaluations is that the A.C.E. failed to correctly identify 4.7 million erroneous enumerations (U.S. Census Bureau 2003, page iv). In addition, it probably mis-identified the residences of large numbers of people in the *P*-sample, leading to both false matches and false non-matches. An extensive program by the Census Bureau of analysis and estimation produced the 1.3 million overcount estimate cited above. However, this program was uniquely tailored to the special circumstances of the 2000 post-census rematching, reinterviewing and duplicate search. Those interested are directed to U.S. Census Bureau (2003).

This paper has described the theory of the DSE, and has discussed how PES in general, and A.C.E. in particular, have implemented that theory. It has described the approximations necessary in real applications and the types of errors that can occur.

It discussed how carefully each of these approximations must be controlled. Obviously, the A.C.E did not successfully measure the large number of duplicates in the 2000 Census. Failure of even extensive probing questions to elicit accurate reports of usual residence was the principal cause. However, the theory and design developed here should be of value in any future coverage measurement program.

ACKNOWLEDGEMENTS

This paper reports the results of research and analysis undertaken by Census Bureau staff. The opinions expressed are those of the author and do not necessarily reflect those of the Census Bureau.

REFERENCES

- BELIN, T.R. (1993). Evaluating sources of variation in record linkage through a factorial experiment. *Survey Methodology*, 19, 13-29.
- CANTWELL, P., and IKEDA, M. (2003). Handling missing data in the 2000 accuracy and coverage evaluation survey. *Survey Methodology*, 29, 2, in press.
- CHILDERS, D. (2001). Accuracy and Coverage Evaluation: The Design Document. DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1 (Revised).
- ESCAP I (2001). Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy. March 1, 2001. (See www.census.gov/dmd/www/pdf/Escap2.pdf)
- ESCAP II (2001). Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy on Adjustment for Non-Redistricting Uses. October 17, 2001. (See www.census.gov/dmd/www/pdf/Recommend2.pdf)
- ESCAP II (2001). Census Person Duplication and the Corresponding A.C.E. Enumeration Status. Executive Steering Committee for A.C.E. Policy II, Report 6.
- FAY, R. (2002). Probabilistic models for detecting census person duplication. *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- FELDPAUSCH, R. (2001). ESCAP II: Census Person Duplication and the Corresponding A.C.E. Enumeration Status. Executive Steering Committee for A.C.E. Policy II, report 6.
- GONZALEZ, M. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section*, American Statistical Association. 73, 7-15.
- GONZALEZ, M., and HOZA, C. (1978). Small-area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*. 73, 361, 7-15.
- GRIFFIN, R. (2000). Accuracy and Coverage Evaluation Survey: Dual System Estimation. DSSD Census 2000 Procedures and Operations Memorandum Series Q-20.
- HAINES, D. (2001). Accuracy and Coverage Evaluation Survey: Computer Specifications for Person Synthetic Estimation (U.S.) Re-issue of Q-30. DSSD Census 2000 Procedures and Operations Memorandum Series Q-46.
- HOGAN, H. (1992). The 1990 Post-Enumeration Survey: An overview. *The American Statistician*. 46, 261-269.
- HOGAN, H. (1993). The Post-Enumeration Survey: Operations and Results. *Journal of American Statistical Association*. 88, 423.
- HOGAN, H. (2001). Accuracy and Coverage Evaluation Survey: Effect of Excluding 'Late Census Adds'. DSSD Census 2000 Procedures and Operations Memorandum Series Q-43. <http://www.census.gov/dmd/www/pdf/Q-43.pdf>
- MARKS, E.S., SELTZER, W. and KROTKI, K.J. (1974). *Population Growth Estimation*. New York: Population Council.
- MARKS, E.S. (1979). The Role of Dual System Estimation in Census Evaluation. In *Recent Developments in PGE*, (K. Krotki). University of Alberta Press. 156-188.
- MULE, T. (2001). ESCAP II: Person Duplication in Census 2000. Executive Steering Committee for A.C.E. Policy II, Report 20.
- MULE, T. (2002). Further Study of Person Duplication Statistical Matching and Modeling Methodology. DSSD A.C.E. Revision II Memorandum Series PP-51.
- NASH, F.F. (2001). ESCAP II: Analysis of Census Imputations. Executive Steering Committee for A.C.E. Policy II, Report 21.

- PETERSEN, C.G.J. (1896). The Yearly Immigration of Young Plaice into the Limfjord from the German Sea. Report of the Danish Biological Station. 6, 1-48.
- RAGLIN, D. (2002). ESCAP II: Effect of Excluding Reinstated Census People from the A.C.E. Person Process. Report 13, <http://www.census.gov/dmd/www/pdf/Report13.PDF>
- ROBINSON, J.G., AHMED, B., DAS GUPTA, P. and WOODROW, K. (1993). Estimates of population coverage in the 1990 united states census based on demographic analysis. *Journal of the American Statistical Association*. 88, 1061-77.
- ROBINSON, J.G. (2001). ESCAP II: Demographic Analysis Results. Executive Steering Committee for A.C.E. Policy II, Report 1.
- SCHENKER, N. (1988). Handling missing data in coverage estimation with application to the 1986 test of adjustment related operations. *Survey Methodology*. 14, 87-97.
- SEKAR, C.C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*. 44, 101-115.
- U.S. CENSUS BUREAU (1985). Evaluating Census of Population and Housing, Statistical Training Document, ISP-TR-5, Washington, D.C.
- U. S. CENSUS BUREAU (2000). Statement on the Feasibility of Using Statistical Methods to Improve the Accuracy of Census 2000.
- U. S. CENSUS BUREAU (2003). Technical Assessment of A.C.E. Revision II, March 12, 2003, <http://www.census.gov/dmd/www/pdf/ACETechAssess.pdf>
- WETROGAN S.I., and CRESCE A.R. (2001). ESCAP II: Characteristics of Census Imputations. Executive Steering Committee for A.C.E. Policy II, Report 22.
- WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*. 81, 338-346.

Handling Missing Data in the 2000 Accuracy and Coverage Evaluation Survey

PATRICK J. CANTWELL and MICHAEL IKEDA¹

ABSTRACT

The Accuracy and Coverage Evaluation survey was conducted to estimate the coverage in the 2000 U.S. Census. After field procedures were completed, several types of missing data had to be addressed to apply dual-system estimation. Some housing units were not interviewed. Two noninterview adjustments were devised from the same set of interviews, one for each of two points in time. In addition, the resident, match, or enumeration status of some respondents was not determined. Methods applied in the past were replaced to accommodate a tighter schedule to compute and verify the estimates. This paper presents the extent of missing data in the survey, describes the procedures applied, comparing them to past and current alternatives, and provides analytical summaries of the procedures, including comparisons of dual-system estimates of population under alternatives. Because the resulting levels of missing data were low, it appears that alternative procedures would not have affected the results substantially. However some changes in the estimates are noted.

KEY WORDS: Cell Imputation; Noninterview Adjustment; Logistic Regression; Dual-System Estimation.

1. INTRODUCTION

Following the 2000 Census in the United States, the Census Bureau conducted the Accuracy and Coverage Evaluation (A.C.E.) survey. The survey had two goals: (1) to measure the level of net undercoverage across the nation and in various demographic and geographic domains through a statistical technique called dual-system estimation, and (2) to produce revised population counts that could be used to adjust for this net undercoverage – if the adjusted numbers were deemed to be more accurate than the initial census counts (Hogan 2003).

In the process of interviewing and following up respondents in the A.C.E., some households were missed, and certain information needed to calculate the dual-system estimates was not collected from other sample respondents. This paper describes the levels of missing data, discusses the procedures used to address the problem in the A.C.E., and provides some results and evaluations. It should be noted that the term “missing data” applies after all follow-up attempts were made in the field. These activities included multiple attempts at interviews, the use of highly trained clerks and technicians to resolve cases, and the follow-up of cases where a second interview could provide additional required information.

The A.C.E. realized three main types of missing data. First, some households were not interviewed because they could not be contacted or the interview was refused. What makes the situation different in the A.C.E. is that to each sample housing unit, *two* noninterview adjustments were applied; one corrected for noninterviews on Census Day, while the other corrected for noninterviews on the day of the A.C.E. interview. As will be shown, the need for two adjustments reflects the different ways that out-movers and in-movers were treated in the dual-system estimation.

The second type of missing data occurred when information for a household or person was available but specific demographic characteristics needed for dual-system estimation were not collected. For missing tenure (owner vs. non-owner), race, and Hispanic origin, a form of nearest-neighbor hot-deck imputation was used to take advantage of the correlations often found among people living in geographic proximity. In general, the values of age and sex are geographically less clustered, but often well predicted by specific conditions, such as the person's relationship (*e.g.*, spouse, child) to the household's reference person, or whether information is available on the person's spouse. Therefore, national donor distributions conditioned on relevant covariates were used to impute for age and sex. Because characteristic imputation for the A.C.E. was similar to that done in the Post-Enumeration Survey following the 1990 Census, the methods and results are not discussed further in this paper.

The third type also involved item missing data. For a small number of people in the A.C.E., not enough information was collected to determine the resident status (whether or not the person was living in the sampled block cluster on Census Day) or the match status (whether or not the person actually matched to someone in the census). Similarly, some people counted in the census lacked sufficient information to determine whether they were correctly enumerated. The status in such cases is said to be “unresolved.” Yet this information is required to compute dual-system estimates. To resolve such cases, a probability of resident (or match or correct enumeration) was assigned as the average weighted value from a set of resolved cases with similar characteristics.

Some of these procedures – described in greater detail below – were applied in similar forms in the 1990 Post-Enumeration Survey, as well as in tests conducted during

¹ Patrick J. Cantwell and Michael Ikeda, Mathematical Statisticians, U.S. Census Bureau, Statistical Research Division, Washington, D.C. 20233-9100.

the 1990s. The main exception is the assignment of a probability in the case of unresolved resident, match, or enumeration status. In the Post-Enumeration Survey and at times for specific tests in the 1990s, these probabilities were computed using a logistic regression model. The method applied in the 2000 A.C.E. used less information than some alternatives such as logistic regression, but was simpler to implement and verify in the tight A.C.E. schedule.

The levels of missing data in the A.C.E. were relatively low, which helped to reduce the potential for additional error in the estimates.

- The household noninterview rates were 3.0% and 1.1% (unweighted), respectively, on Census Day and Interview Day.
- The imputation rates for the five A.C.E. characteristics required for dual-system estimation ranged from 1.4% to 2.5% (unweighted and weighted).
- Among people in the A.C.E., the rates of unresolved resident and match status were 2.3% and 1.2% (unweighted), respectively; among census enumerations, only 3.0% (2.6% weighted) of the sample had unresolved enumeration status.

When assigning probabilities for unresolved status, the success of the variables used to define imputation cells was mixed. Variables that used information related to an individual's processing in the survey operations discriminated well among cells. However, variables describing the person's demographic characteristics appear to have been generally less successful.

Section 2 contains background information about the A.C.E. and dual-system estimation. The A.C.E. non-interview adjustment is discussed in section 3. For persons with unresolved resident, match, or enumeration status, a probability was assigned according to procedures described in section 4. Section 5 examines the effect of some alternatives to the A.C.E. missing data procedures on the dual-system estimates of the population. Finally, a few observations are recounted in section 6. For a detailed description of the missing data procedures for the 2000 A.C.E., see Cantwell (2001). Summaries of missing data can be found in Cantwell *et al.* (2001).

In what follows, unweighted frequencies and proportions are generally given. Unless noted otherwise, the weighted numbers are very close. However, the probabilities assigned to unresolved cases in Tables 4, 5, and 6 are the actual weighted ones used in the estimation.

2. A BRIEF ACCOUNT OF THE SURVEY AND DUAL-SYSTEM ESTIMATION

Through the Accuracy and Coverage Evaluation (A.C.E.), the Census Bureau attempted to measure and adjust for the historical differential net undercount observed

in the U.S. Census (Anderson and Fienberg 1999, page 29). Like the 2000 Census, the A.C.E. covered the entire nation. (A separate sample and analysis were conducted for Puerto Rico.) A sample of about 300,000 housing units in 11,303 block clusters was selected (Fenstermaker 2000, Childers 2000).

To estimate coverage of the population, the A.C.E. relied on dual-system estimation, a method based on capture-recapture methodology (Peterson 1896, Sekar and Deming 1949). Suppose one considers only those housing units contained in the sample of block clusters selected for the A.C.E. After the census enumeration – but without using *any* information collected in the census – the Census Bureau independently interviewed people in the A.C.E. sample and obtained a roster of people living in the units on Census Day, April 1, 2000. These results were then matched to (compared with) the census enumeration in those block clusters to estimate how many people were missed. Within the sample block clusters, the units enumerated independently in the A.C.E. were defined as the *P*-Sample, and those enumerated in the census as the *E*-Sample.

In the same sample of block clusters, comparisons and analyses were made to estimate the proportion of census enumerations that were correct, that is, complete, unique, and recorded in the proper location. Erroneous enumerations include people who are duplicated or fictitious, or should not be counted at that address, for example, because their usual residence was elsewhere, such as in a college dormitory. The resulting dual-system estimator is

$$\hat{N} = (C - I) \hat{p}_{ce} \left(\frac{1}{\hat{p}_{match}} \right), \quad (1)$$

where C is the official census count, including imputed persons and erroneous enumerations; I is the number of whole-person imputations; \hat{p}_{ce} is the weighted estimate of the proportion of correct enumerations in the census; and \hat{p}_{match} is the weighted estimate of the proportion of *P*-Sample people who match to someone enumerated in the census. People are imputed, for example, when a census enumerator confirms that a certain number of people live at an eligible address, but sufficient additional information cannot be gathered. The actual number of whole-person imputations is known and can be removed from C in the estimate.

Dual-system estimates were calculated separately within population subgroups called post-strata. Post-stratum estimates were then used to determine adjustment factors to be applied to all people counted in the census according to their specific post-stratum. Finally, adjusted counts for any geographic area were calculated by summing the adjusted counts across post-strata in the area. For more detailed information on A.C.E. field operations and dual-system estimation in general, see Childers (2000) and Hogan (1993, 2003), respectively.

3. NONINTERVIEW ADJUSTMENT

Noninterview adjustment was performed only on the *P*-Sample; in the census (and, thus, in the *E*-Sample), people in all known housing units were accounted for through a variety of procedures. The small number of housing units whose information was collected by a proxy respondent, often a neighbor or building manager, were treated as valid interviews and are not the subject of the noninterview adjustment. Because people moved in and out of housing units between Census Day and the time of the A.C.E. interview, the Census Bureau had to consider the mover status – out-mover, in-mover, or non-mover – of all people in the *P*-Sample, as well as the interview situation at the two different moments. Out-movers were living in the housing unit in question on Census Day, but had moved out before Interview Day. The situation was reversed for in-movers. Non-movers lived in the unit on both days. At the time of the A.C.E. interview, in *one interview* questions were asked to determine who lived in the household on Interview Day and who lived there on Census Day. Mover status was assigned to each person in the sample, and two rosters were created for each household – the Census Day roster and the Interview Day roster.

The A.C.E. used in-movers to estimate the *number* of *P*-Sample movers, while using out-movers to estimate the *match rate* of the movers. The weighted *P*-Sample total, that is, the denominator of \hat{p}_{match} in equation (2), is estimated as the weighted total of all non-movers and in-movers. Yet the weighted number of *P*-Sample matches is estimated by adding the number of matches among non-movers to the product of the number of in-movers and the match rate for out-movers:

$$\hat{p}_{\text{match}} = \frac{M_{nm} + N_{im} \times \frac{M_{om}}{N_{om}}}{N_{nm} + N_{im}}, \quad (2)$$

where *N* (people) and *M* (matches) are indexed by *nm*, *im*, and *om*, representing non-movers, in-movers, and out-movers, respectively. All in-movers and non-movers were generally assumed to be A.C.E. Interview Day residents. (People living in group quarters, such as college students in dormitories, were not eligible for the *P*-Sample.)

The mover procedure used in the A.C.E. differed from that used in the 1990 Post-Enumeration Survey. In 1990 in-movers were used to estimate the number of movers and their match rate. For the latter, the in-movers had to be matched back to their address on Census Day. That procedure was changed for the census tests conducted during the 1990's to accommodate the planned use of sampling for census nonrespondents. When the U.S. Supreme Court ruled against the sampling plan in 1999 (*Department of Commerce v. United States House of Representatives*, 525 U.S. 316, 1999), it was thought that

changing the mover procedure again so late before the census would introduce unacceptable risks.

Due to the mover procedure described above, each housing unit had two interview statuses – one based on the housing unit's situation as of Census Day, and the other based on the day of the A.C.E. interview. A unit that was vacant, removed from the list of eligible housing units (because, for example, it was demolished or used only as a business), or in certain special places was not considered an interview or a noninterview. Table 1 provides a fictional but illustrative block cluster. It demonstrates how the status of a housing unit on Census Day and Interview Day would have been determined.

Results of the A.C.E. interviewing operation are shown in Table 2. Of the 261,969 housing units occupied on Census Day, 7,794 (3.0 percent) were noninterviews. The corresponding numbers for Interview Day were 267,155 and 3,052 (1.1 percent).

As different interview statuses were possible for a housing unit on Census Day and Interview Day, different noninterview adjustments were required for each day. Each of the two adjustments generally spread the weights of noninterviewed units over interviewed units in the same noninterview cell: the sample block cluster crossed with the type of basic address, defined as single-family, multi-unit (such as apartments and condominiums), or all others. Other characteristics, known for all housing units, could have been used to define the cells. However, the cells were defined to take advantage of the typical local homogeneity, and of the fact that people living in, for example, apartments share many of the characteristics – household size, propensity to move, *etc.* – that are related to capture probabilities in the census.

The noninterview adjustment based on the Census Day status of housing units was used to adjust the person weights of non-movers and out-movers. Similarly, the Interview Day noninterview adjustment was used to adjust the person weights of in-movers. Within a noninterview cell, the adjustment factor for *Census Day* was computed as the weighted sum of interviews and noninterviews for Census Day divided by the weighted sum of interviews for Census Day. A housing unit's weight was the inverse of the final selection probability of its block cluster into the A.C.E. sample. (These weights were trimmed in a very small number of clusters.)

The noninterview adjustment factor for Interview Day was computed as above, but with its status – interview, noninterview, vacant, or delete – being considered for Interview Day rather than for Census Day. The example in Table 1 demonstrates the calculation of the noninterview adjustment for the fictional block cluster. Because the noninterview rates were so small, the noninterview adjustment factors were close to 1 for most housing units in the sample. For Census Day, the factors were less than 1.10 for more than 92% of the units; for Interview Day, the factors were less than 1.10 for over 98% of the units.

Table 1
An Example of Adjustment for Noninterviews

Consider a block cluster with nine housing units, all having the same type of basic address, for example, all single family homes, as depicted below					
Housing Unit	Weight	Actual Situation	Status of (and Information from) A.C.E. Interview	Census Day Status	A.C.E. Interview Day Status
1	100	Resident on 4/1/00 and at time of A.C.E. interview	Interviewed in A.C.E.	Interview	Interview
2	100	Resident on 4/1 and at time of A.C.E. interview	Neighbor (proxy) interviewed in A.C.E.	Interview	Interview
3	100	Resident on 4/1 and at time of A.C.E. interview	No one interviewed in A.C.E.	Noninterview	Noninterview
4	100	Vacant on 4/1, resident at time of A.C.E. interview	Interviewed in A.C.E., knows of 4/1 status	Vacant	Interview
5	100	Vacant on 4/1, resident at time of A.C.E. interview	Interviewed in A.C.E., no knowledge of 4/1 status	Noninterview	Interview
6	100	Vacant on 4/1, resident at time of A.C.E. interview	No one interviewed in A.C.E.	Noninterview	Noninterview
7	100	Resident on 4/1, vacant at time of A.C.E. interview	Information obtained from proxy	Interview	Vacant
8	100	Resident on 4/1, vacant at time of A.C.E. interview	No info on 4/1 status; Census staff determines vacant at time of A.C.E.	Noninterview	Vacant
9	100	Resident on 4/1, different resident at time of A.C.E. interview	Interviewed in A.C.E., knows of 4/1 status	Interview	Interview

In this noninterview cell (sample block cluster \times type of basic address), people in interviewed housing units would have received the following noninterview adjustments:

- (1) to the person weights of non-movers and out-movers, the Census Day noninterview adjustment = $800 / 400 = 2.0$
- (2) to the person weights of in-movers, the A.C.E. Interview Day noninterview adjustment = $700 / 500 = 1.4$

Table 2
Status of Household Interviews in the A.C.E. (Unweighted)

	Census Day		A.C.E. Interview Day	
	Number	Percent	Number	Percent
Total Housing Units	300,913	100.0	300,913	100.0
Interviews	254,175	84.5	264,103	87.8
Noninterviews	7,794	2.6	3,052	1.0
Vacant Units	28,472	9.5	29,662	9.9
Deleted Units	10,472	3.5	4,096	1.4
Noninterview rate ¹	3.0%		1.1%	

¹ Noninterview rate = Noninterviews / (Interviews + Noninterviews)

When the unweighted number of noninterviewed units in a given noninterview cell was more than twice the unweighted number of interviewed units, the weights of the noninterviewed units in this cell were spread over the interviewed units in a broader set of noninterview cells. This remedy was needed for only 65 cells for the Census Day adjustment, and 13 cells for the Interview Day adjustment. The prescribed procedure differs from the usual collapsing of sparse cells, but allowed us to address such cells in a simple automated fashion. This capability was important under a very tight schedule when it was impossible to predict which cells would have too few

interviews. For evaluation purposes, the housing-unit weights were later re-computed under a collapsing scheme, and compared to the weights as determined in the A.C.E. Again, due to the low rates of noninterview, the weights were the same for most units, and close for the rest. The effect on the resulting dual-system estimates is shown in section 5.2.

4. ASSIGNING PROBABILITIES FOR UNRESOLVED CASES

After all A.C.E. follow-up activities were completed, there remained a small fraction of the A.C.E. sample without enough information to compute the components of the dual-system estimator given in equation (1). Their status was said to be "unresolved."

4.1 Unresolved Cases and Their Frequencies

One component of the dual-system estimator in equation (1) is \hat{p}_{match} , the estimated proportion of the P -Sample who match to someone enumerated in the census. In (2) for \hat{p}_{match} , when estimating the number of people (N_{nm} , N_{om}) or matches (M_{nm} , M_{om}) among non-movers and out-movers, only Census Day residents of the sample block clusters were considered; someone who usually lives in a nursing home, for example, was omitted from the computation.

Thus, for each person in the *P*-Sample, determining resident status and match status was required.

After follow-up operations were completed, all people in the *P*-Sample who were eligible to be matched to the census were classified into three types according to their status as a resident in their sampled block on *Census Day*: residents, nonresidents, and unresolved persons – those for whom there was not enough information to determine the resident status. Further, each confirmed or possible (unresolved) Census Day resident in the *P*-Sample was determined to be a match, a nonmatch, or unresolved match. The match status for confirmed Census Day nonresidents, such as in-movers, was not used in the estimation. The estimator in (1) also requires an estimate of the proportion of correct enumerations in the census, \hat{p}_{ce} . After whole-person imputations were removed from the *E*-Sample, each remaining person had one of three types of enumeration status: correct, erroneous, or unresolved.

Table 3 summarizes the frequencies of resident and match status in the *P*-Sample, and enumeration status in the *E*-Sample. The table also shows results for non-movers and out-movers in the *P*-Sample. One can see that the extent of unresolved cases is quite small: 2.3% for resident status, 1.2% for match status, and 3.0% for enumeration status. (The weighted rates are 2.2%, 1.2%, and 2.6%, respectively.) In the 1990 Post-Enumeration Survey, the rate of unresolved matches was 1.9%, and unresolved enumerations was 2.4%. (Resident status was not defined in a manner comparable to 2000.) Care must be taken, however, as the definitions of the several statuses were slightly different in 1990.

4.2 Assigning Probabilities to Unresolved Cases

In the A.C.E., a form of cell imputation was used to assign probabilities for sample cases with unresolved resident, match, or enumeration status. All people in the sample – resolved and unresolved – were placed into groups called imputation cells based on operational and demographic characteristics. Different variables were used to define cells for each type of status. Within each imputation cell the weighted average of 1's and 0's (representing, *e.g.*, match and non-match, respectively) among the resolved cases was calculated, and that average was imputed for all unresolved persons in the cell. More details are provided below.

In the 1990 Post-Enumeration Survey, hierarchical logistic regression was used to calculate probabilities of match and correct enumeration for cases with missing information. (Due to the procedure used to treat movers in 1990, resident status played a different role then.) The model and some results are discussed in Belin *et al.* (1993).

During the 1990s, the Census Bureau originally planned to produce in 2000 adjusted census estimates for each of the 50 states (and the District of Columbia) using data collected only from that state. This approach affected the strategy for treating unresolved status in two ways. First, within each state, there would be far fewer data – resolved cases – on which to build a logistic regression model. Second, there would be 153 different models to examine and verify, separate models for resident, match, and enumeration status in each state. Because the production schedule for the A.C.E. provided only about three weeks for addressing all

Table 3
Final Status Frequencies for the *P* and *E*-Samples (Unweighted)

<i>P</i> -Sample	Total people ¹	Final resident status			Resident rate for resolved cases
		Confirmed resident	Confirmed nonresident	Unresolved resident	
U.S. Total	653,337	95.8%	1.9%	2.3%	98.1%
Mover status					
Non-mover	627,992	96.6%	1.7%	1.7%	98.3%
Out-mover	25,345	75.2%	7.5%	17.4%	91.0%
<i>P</i> -Sample	Total people ²	Final match status			Match rate for resolved cases
		Match	Nonmatch	Unresolved match	
U.S. Total	640,945	90.3%	8.5%	1.2%	91.4%
Mover status					
Non-mover	617,490	91.1%	8.0%	0.9%	91.9%
Out-mover	23,455	67.8%	21.7%	10.5%	75.8%
<i>E</i> -Sample	Total people	Final enumeration status			Correct enumeration rate for resolved cases
		Correct enumeration	Erroneous enumeration	Unresolved enumeration	
U.S. Total	704,602	92.6%	4.4%	3.0%	95.5%

¹ Those in the *P*-Sample eligible to be matched to the census.

² Confirmed or possible residents in the *P*-Sample.

aspects of missing data, it was believed that a procedure to handle unresolved status that was simpler to implement and verify would reduce the risk of not completing the dual-system estimates under the imposed deadline. Cell imputation provided the desired simplicity, but its accuracy relative to that under logistic regression modeling had to be evaluated in subsequent testing.

During census tests in 1995 and 1996, certain types of unresolved status were addressed using logistic regression, while cell imputation was used for other types. The latter procedure was used exclusively in the Census Dress Rehearsal in 1998 (Ikeda, Kearney and Petroni 1998), when the Census Bureau was still preparing to produce estimates independently within each state. Data from these tests indicated that the exact method of calculating probabilities for unresolved status (match, resident, or correct enumeration) had only a minor effect on the dual-system estimates. Details of this research can be found in Petroni (1997, 1998a, 1998b, and 1998c).

With the decision by the U.S. Supreme Court in 1999 (*Dept. of Commerce v. U.S. House of Representatives*), the Census Bureau changed the design of the survey and removed the restriction that adjusted estimates be based solely on data from within each state. However, there remained concerns about implementing a logistic regression approach that had not been tested in the Dress Rehearsal. Further, there was no guarantee that available software would adequately run logistic models on data sets the size of the entire A.C.E. sample (between 640,000 and 750,000 people). Based on these concerns and research findings on relative accuracy, a decision was made to use the simpler procedure, cell imputation, to resolve missing status in the A.C.E.

To demonstrate how cell imputation was applied in the A.C.E., one can look at resident status; the method was

applied analogously to match and enumeration status. First, all non-movers and out-movers in the *P* Sample were placed into a number of imputation cells according to operational and demographic characteristics, as defined in Table 4; in-movers were left out, as their Census Day resident probability was 0 by definition. Among the resolved cases in cell *i*, denoted by the set *R*(*i*), an indicator variable for resident status was defined as $1_{res,j} = 1$, if person *j* was a resident in the household on Census Day, or 0, otherwise. Then within cell *i*, the weighted proportion of Census Day residents, was computed:

$$P(res)_i = \frac{\sum_{j \in R(i)} w_j 1_{res,j}}{\sum_{j \in R(i)} w_j} \quad (3)$$

where w_j is the weight of person *j* incorporating all stages of sampling. $P(res)_i$ was then assigned to each unresolved person in cell *i*, that is, each of the 15,082 people (2.3% of 653,337) with unresolved resident status. (The exception is for match code group 7, as explained below.) Table 4 provides the resident probabilities assigned within the cells. This assignment defines for all cases – resolved and unresolved – an “extended” indicator, allowing values between 0 and 1:

$$1'_{res,j} = \begin{cases} 1_{res,j}, & \text{if } j \in R(i) \\ P(res)_i, & \text{otherwise} \end{cases} \quad (4)$$

The estimated numbers of non-movers and out-movers in the *P*-Sample in (2), N_{nm} and N_{om} , respectively, are then computed by attaching the person weight and summing the indicator $1'_{res,j}$ over the non-movers and out-movers, respectively, in all cells. The number of matches, M_{nm} or

Table 4
Imputation Cells for Resolving Resident Status in the *P*-Sample

<i>P</i> Sample Match Code Group	Owner		Non-Owner	
	Nonhispanic White	Others	Nonhispanic White	Others
1. Matches needing follow-up	0.982	0.986	0.993	0.991
2. Possible matches	0.973	0.968	0.966	0.972
3a. Partial household nonmatches needing follow-up; aged 18-29, child of reference person	0.755	0.901	0.883	0.928
3b. Partial household nonmatches needing follow-up; others not in 3a	0.956	0.971	0.959	0.969
4. Whole household nonmatches needing follow-up, not conflicting households	0.920	0.943	0.911	0.914
5. Nonmatches from conflicting households	0.910	0.927	0.945	0.954
6. Resolved before follow-up	0.993	0.990	0.990	0.988
7. Insufficient information for matching (Weighted column average of groups 1-5 and 8)	0.813	0.867	0.844	0.872
8. Potentially fictitious or said to be living elsewhere on Census Day	0.119	0.123	0.177	0.157

M_{om} , and thus, \hat{p}_{match} , are determined analogously, as is \hat{p}_{ce} , in the case of enumeration status.

In the Census Dress Rehearsal of 1998, cell imputation for unresolved resident probability was used with only three cells: persons sent to follow-up, persons not needing follow-up, and persons with insufficient information for matching. For the third cell, which contained no resolved cases, a proportion based on all resolved cases in the first two cells was assigned. Results from the Dress Rehearsal (Kearney and Ikeda 1999) suggested that dividing the *P*-Sample into the various match code groups would be helpful. Further research and discussion suggested adding other demographic variables within match code group. The larger A.C.E. sample size in 2000 made it possible to support a larger set of imputation cells.

For the A.C.E. in 2000, match code groups 1 through 7 were determined from the match codes and other variables derived *before* the follow-up operation, as explained in Childers (2000). Group 8 was formed differently. Some information from the follow-up operation was coded in time for the A.C.E. missing data procedures. (Under the original schedule, this information would have become available too late to be of use.) *After* the follow-up operation, a small number of people in the *P*-Sample were coded as being potentially fictitious or said to be living elsewhere on Census Day. Among the resolved cases in this group, the probability of being a resident was much lower than for resolved people in other groups. Thus, people satisfying the conditions for group 8 were placed there first, and each of the remaining people was placed appropriately in one of the first seven groups.

Two tenure categories were used: owners and non-owners. Persons were also placed into one of two race-ethnicity categories: Nonhispanic white, and all others. People of multiple races (for example, a person responding as White and Asian) were placed in the latter group. Match code group 3, partial household nonmatches, was split into two subgroups. The first, 3a, included persons in group 3 who were 18 to 29 years of age and were listed on the A.C.E. household roster as a child of the reference person. These were young people many of whom were attending college, sharing residence with colleagues, or moving in and out of their parents' residence. Classification and regression tree analysis, applied to data from the Census Dress Rehearsal of 1998, suggested that this combination of characteristics would discriminate well with respect to resident status. The group 3b included all other persons in group 3.

The resident probability for unresolved *P*-Sample persons was computed as described above, except for those in match code group 7 – people with insufficient information for matching. Within this row in Table 4, there were essentially no resolved cases from which to extract a probability of being a Census Day resident. Because of their lack of information – most of these cases did not even have

a valid name – these people did not go through the matching operation and were not sent to follow-up. To determine a resident probability for these cases, a weighted proportion of Census Day residents (1's and 0's) was computed among the resolved cases in match code groups 1 through 5 and 8, separately for each of the four tenure \times race/ethnicity classes. This probability was then assigned to those in group 7. Left out of this computation were those people who were resolved before follow-up (group 6). Observations from the Dress Rehearsal indicated that, in terms of their demographic and operational characteristics, people in group 7 tended to be more like those in groups 1–5 and 8, than like those in group 6.

The issue of unresolved matches was treated like that for unresolved resident status in (3) and (4), with resident status replaced by match status, but with a different set of cells, as is seen in Table 5. Confirmed nonresidents were excluded from the computations of match probabilities.

For unresolved match probability in the Dress Rehearsal, only one imputation cell was used within each of the geographic sites. Subsequent analysis (Kearney and Ikeda 1999) showed that mover status (non-mover vs. out-mover) discriminated well between matches and nonmatches among the resolved cases. Thus, for the 2000 A.C.E. mover status was used to define imputation cells for match status. The housing-unit address match code refers to the initial match between housing units on the independent (A.C.E.) listing and the census address list; conflicting housing units, determined during A.C.E. person match activities, were those where the census and A.C.E. rosters had two completely different lists of residents for Census Day (Childers 2000).

It should be noted that 98.3% of the unresolved matches (7,693 of 7,826) were people with insufficient information for matching. As mentioned above, most of them did not have a valid name, and almost all (7,506) were not sent to follow-up. Further, their rate of missing characteristics was much higher than average. Therefore, little useful predictive information was available when forming imputation cells for match status. Variables such as age and ethnicity – that had a higher chance of being imputed and might be of questionable quality – were avoided.

People with at least one imputed demographic variable (age, sex, tenure, race, or Hispanic origin) were grouped when resolving match status. Unpublished studies indicated that – at least among resolved cases in the Dress Rehearsal – the presence of these imputed characteristics is negatively associated with the propensity to be a match. Out-movers from a unit that was a nonmatch or a conflicting household were not separated according to this variable to ensure a reasonable number of resolved cases in each cell from which to estimate the proportion of matches.

In the *E*-Sample, unresolved enumeration status was addressed as discussed above. See Table 6.

Table 5
Imputation Cells for Resolving Match Status in the *P*-Sample

Mover Status		Housing-Unit Address Match Code		
		Housing unit was a match		Housing unit was a nonmatch or the household was conflicting
Non-mover	No imputed characteristics ¹ 0.945	1 or more imputed characteristics 0.901	No imputed characteristics 0.690	1 or more imputed characteristics 0.567
Out-mover	No imputed characteristics 0.798	1 or more imputed characteristics 0.791		0.516

¹ Among the characteristics age, sex, tenure, race, or Hispanic origin.

Table 6
Imputation Cells for Resolving Enumeration Status in the *E*-Sample

E-Sample Match Code Group		No Imputed Characteristics ¹		1 or More Imputed Characteristics
1.	Matches needing follow-up		0.977	0.977
2.	Possible matches		0.968	0.968
3a.	Partial household nonmatches; aged 18-29, child of reference person		0.871	0.908
3b.	Partial household nonmatches; others not in 3a		0.974	0.960
4.	Whole household nonmatches where the housing unit matched; not conflicting households	Nonhispanic White 0.965	Others 0.974	0.958
5.	Nonmatches from conflicting households; for housing units not in regular nonresponse follow-up		0.975	0.965
6.	Nonmatches from conflicting households; housing units in regular nonresponse follow-up		0.914	0.926
7.	Whole household nonmatches, where the housing did not match in housing-unit matching	Nonhispanic White 0.959	Others 0.947	0.950
8.	Resolved before follow-up	Nonhispanic White 0.995	Others 0.990	0.979
9.	Insufficient information for matching		0 (assigned by definition)	
10.	Targeted extended search cases ²		0.928	0.858
11.	Potentially fictitious people		0.058	0.088
12.	People said to be living elsewhere on Census Day		0.229	0.210

¹ Among the characteristics age, sex, tenure, race, or Hispanic origin.

² Targeted extended search refers to a field operation conducted to reduce the variance in the dual-system estimates caused by clustered geocoding errors. For more information, see Navarro and Olson (2001).

As with resident status for *P*-Sample people, a key factor in determining enumeration status was the *E*-Sample person's match code group, although the match code groups were defined differently for the two samples. Similar to the *P*-Sample, people coded as potentially fictitious or said to be living elsewhere on Census Day during the follow-up operation were first placed in groups 11 or 12, respectively. The remainder of the *E*-Sample was then placed in the appropriate match code group, as defined in the table. Group 3 was split into two subgroups, as when determining

resident status in the *P*-Sample. That is, people aged 18 to 29 who were children of the reference person were separated. Other characteristics used to define cells were the presence or absence of imputed characteristics, as defined in the imputation cells for match status; and whether the person was Nonhispanic white or any other race-ethnicity combination. It should be noted that, according to A.C.E. procedures, anyone in the *E*-Sample with insufficient information for matching (group 9) was automatically assigned an enumeration probability of 0.

4.3 Comparing Probabilities Under Cell Imputation and Logistic Regression

It can be insightful to compare the probabilities assigned to cases with unresolved status under alternative procedures. Belin (2001) presents such a comparison under a logistic regression model that considered 186 predictors for resident and match status, and 202 predictors for enumeration status. The variables included most of those used in the cell estimation described in section 4.2, as well as individual demographic characteristics, such as age, gender, and relationship to the household's reference person; information about the A.C.E. interview, such as whether the respondent was a proxy; information derived from the sampling process; local-area features, such as whether the area was urban or non-urban; and the interactions among the variables. As the models were fit to the resolved cases

sent to follow-up, and then applied to unresolved cases to predict a probability, the models are ignorable in the sense that unresolved status is not considered as a covariate in the underlying model. (See Rubin 1976.)

Tables 7 and 8 summarize the probabilities assigned to unresolved cases under A.C.E. cell imputation and the logistic modeling averaged over the different match code groups. Recall that cell imputation probabilities were computed from weighted data as in (3), while the logistic regression models were run on unweighted data. The predicted probabilities for the two procedures were averaged across all unresolved people unweighted. With an exception to be mentioned later, probabilities and estimates in the A.C.E. were typically similar when using unweighted and weighted data, as the sample was designed to avoid a wide range of weights.

Table 7
Average Resident and Match Probabilities Under Cell Imputation and Logistic Regression

P-Sample Match Code Group	Resident Status			Match Status		
	Avg. Probability Assigned			Avg. Probability Assigned		
	Number Unresolved	Cell Imputation	Logistic Regression	Number Unresolved	Cell Imputation	Logistic Regression
1. Matches needing follow-up	767	0.989	0.983	4	0.848	0.941
2. Possible matches	352	0.970	0.962	131	0.889	0.837
3. Partial household nonmatches	1,306	0.956	0.951	71	0.893	0.050
4. Whole household nonmatches	1,610	0.917	0.926	36	0.770	0.010
5. Nonmatches, conflicting household	1,455	0.940	0.927	49	0.616	0.070
6. Resolved before follow-up	129	0.990	0.990	23	0.842	0.940
7. Insufficient information	7,506	0.844	0.851	7,506	0.835	0.880
8. Fictitious, living elsewhere	2,402	0.148	0.167	6	0.655	0.041

Table 8
Average Enumeration Probabilities Under Cell Imputation and Logistic Regression

E-Sample Match Code Group	Enumeration Status		
	Avg. Probability Assigned		
	Number Unresolved	Cell Imputation	Logistic Regression
1. Matches needing follow-up	711	0.977	0.986
2. Possible matches	305	0.968	0.967
3. Partial household nonmatches	2,191	0.962	0.963
4. Whole household nonmatches where the housing unit matched; not conflicting	4,813	0.967	0.974
5. Nonmatches from conflicting households; housing units <u>not</u> in nonresponse follow-up	532	0.973	0.973
6. Nonmatches from conflicting households; housing units in nonresponse follow-up	779	0.917	0.926
7. Whole household nonmatches, where the housing unit did not match	3,881	0.954	0.961
8. Resolved before follow-up	179	0.990	0.982
9. Insufficient information for matching	0	----	----
10. Targeted extended search cases	2,902	0.918	0.679
11. Potentially fictitious people	1,690	0.064	0.077
12. People said to be living elsewhere on Census Day	3,152	0.225	0.280

Comparing procedures, one sees almost no difference in the average probabilities assigned for resident status. This is not surprising, as cell imputation used the match code group (among other variables) to define cells. Match status presents a different story. To recall, match code group was not used in the cell imputation, as almost all unresolved matches (98.3% of 7,826; 7,506 before the follow-up operation, and 187 more after follow-up) had insufficient information for matching. The first two groups have slightly different probabilities assigned under the two procedures. But in groups 3, 4, and 5, all nonmatches before follow-up, the average probabilities are high under cell imputation (0.893, 0.770, and 0.616), and very low under logistic regression (0.050, 0.010, and 0.070). Of the 156 cases in the three cells, 134 were people each of whom was given an initial code indicating a "nonmatch"; later it was determined correctly that the person had insufficient information for matching. In almost every case, the A.C.E. interviewer recorded a name like "Child Jones", "José Don't Know", or "Unknown Smith". Such cases should have been caught before matching by a clerk, and assigned an initial code of insufficient information. Instead, a match to the census was attempted and failed. If not for this error, such people would have been placed in group 7, where their match probability under logistic regression would have been much higher. Thus, for this small set of 134 cases, the logistic variable, match code group, takes on an incorrect value, and the model predicts a probability – much too low – based on the many resolved cases in group 3, 4, or 5 *who really were nonmatches*, but were sent to follow-up primarily to resolve their resident status, not their match status.

The predicted match probabilities in group 8 were also very different. However, with only six unresolved cases, the effect on estimation would be minimal.

Comparing average enumeration probabilities by match code group in Table 8, one sees almost no difference except in group 10, targeted extended search cases. There, the average probability assigned by cell imputation, 0.918, is much higher than that predicted by logistic regression, 0.679. The difference can be explained by the weighting. In the *E*-Sample, of 32,334 people eligible for the targeted extended search operation, 8,298 (all in match code group 10) were sampled out to contain costs and given an A.C.E. weight of 0. The matching operation did not try to determine whether the 8,298 cases were enumerated correctly or not, but simply left them on the data file as erroneous enumerations. Probabilities based on cell imputation were assigned as in equations (3) and (4), incorporating the A.C.E. weight. This removed from the computation those who were sampled out of the A.C.E. The logistic regression model was run on unweighted data and included the 8,298 cases in group 10, bringing down the probability of a correct enumeration predicted for the 2,902 people with unresolved enumeration status.

5. THE EFFECT OF SOME ALTERNATIVE MISSING DATA PROCEDURES ON DUAL-SYSTEM ESTIMATES

In the last section, predicted probabilities were compared across two options for treating cases with unresolved status. But the ultimate effect of competing procedures is seen in the resulting dual-system estimates. In this section, several alternatives to those used in the A.C.E. for addressing missing data are compared via the resulting estimates. When they differ significantly, it is not clear which procedure is to be preferred. It should be noted that the A.C.E. estimates released by the U.S. Census Bureau in March of 2001 have been revised following further analyses (Haines 2003). Even though the A.C.E. data are flawed and A.C.E. estimates should generally not be used, it is believed that they are adequate to evaluate the differences in the estimates caused by alternative missing data approaches.

5.1 Results from an Early Evaluation

In the months after initial dual-system estimates from the A.C.E. were released, alternatives to the applied missing data procedures were studied. There were several reasons: estimating the variation that might result from the alternatives, incorporating this variation into total error and loss function analysis for the A.C.E. dual-system estimates, and investigating the viability of non-ignorable missingness procedures for addressing unresolved status. As the results are available in Keathley, Kearney, and Bell (2001), only a summary will be provided here.

Three alternatives involving the noninterview adjustment were examined. The first defined cells differently for the adjustment, adding variables such as race, Hispanic origin, tenure, and household size, as determined from a match to the census file. This procedure tended to produce larger dual-system estimates. Two other noninterview alternatives had no apparent effect on the estimates. In one, a nearest-neighbor noninterview adjustment, the weight of a non-interviewed household was added to that of the nearest interviewed household in the sorted file. In the second, the last 30% of A.C.E. interviews completed were labeled as "late" interviews. The weights of noninterviewed units were added only to the weights of late interviews. These alternatives tried to take advantage of the anticipated homogeneity of units induced by geographic proximity or time of response to the A.C.E.

The other alternatives described in Keathley *et al.* (2001) address unresolved resident, match, or enumeration status. A "late" data approach used information collected only from the last 30% of interviews in the *P*-Sample, or housing units that required nonresponse follow-up in the *E*-Sample. By itself, this approach did not appear to affect the dual-system estimates. The remaining alternatives involved logistic regression models to predict probabilities for

unresolved cases. First, an ignorable logistic model, the one described above (Section 4.3) in Belin (2001), was applied to unresolved resident, match, and enumeration status and tended to produce smaller dual-system estimates (47,481 smaller for the U.S. total). However, it appears that the lowered (on average) enumeration probabilities assigned to the 2902 unresolved cases in the *E*-Sample match code group 10 (see section 4.3) would have more than accounted for this decrease.

Perhaps more interesting are three alternatives that attempted to construct non-ignorable logistic models by lowering the probabilities assigned to unresolved cases, on the premise that ignorable models may overstate the underlying probabilities (Belin 2001). Data from the 1990 Post-Enumeration Survey and its evaluation follow-up were used to estimate non-ignorable effects and incorporate them into the 2000 logistic models. This strategy tended to produce larger dual-system estimates when applied to unresolved match probabilities, and smaller estimates when applied to resident or enumeration probabilities. This result is not surprising, based on equation (1) and the fact that the average match probability assigned to cases with unresolved resident status is less than that for cases with resolved resident status. Although the study's authors conclude that "[t]here is no evidence to suggest that the non-ignorable missingness procedures that we considered are or are not viable alternative missing data procedures" (Keathley *et al.* 2001, page 2), Belin's approach takes a promising step toward addressing the non-ignorability of the missing status.

5.2 Analyses on Other Alternative Procedures

In this section, differences in the dual-system estimates are presented under six numbered alternatives described and motivated below. The results are provided in Table 9 for the U.S. total and for breakdowns by race-ethnicity, tenure, and age. For a precise definition of the race-ethnicity domains, see Kostanich (2001). (Note that a small part of the U.S. population was not part of the A.C.E. universe.) For each alternative, the three numbers given are (a) the difference: the alternative estimate minus the A.C.E. estimate; (b) the standard error of that difference; and (c) the percent relative difference.

Alternative (1) reconsidered the noninterview procedure as applied in the A.C.E. to adjustment cells with a relatively small number of completed interviews. (See section 3.) In this alternative, instead of spreading weights from non-interviewed units over a wider range of cells, cells with too few interviews were collapsed with nearby cells, and noninterview adjustment factors were computed afresh in the newly created cells. Except for Nonhispanic Blacks,

none of the estimated differences in Table 9 under this alternative are statistically significant (greater than two standard errors). Similarly, except for several race-ethnicity domains less than two million in size, none of the relative differences are greater than 0.01%.

Alternatives (2), (3), and (4) were derived after examining the effects of the variables used in the imputation cells on the resulting assigned probabilities. From the probabilities assigned in Tables 4 and 6, it is clear that the match code groups discriminated well with regard to resident and enumeration status. Yet it appears that dividing the cells based on demographic variables, such as "Nonhispanic white" vs. "Other," made less of a difference. To investigate the effect of demographic variables on the imputation, new probabilities were assigned for unresolved status without using them. Specifically, all resolved and unresolved cases were combined across cells for Nonhispanic white and Other (resident and enumeration status), for match code groups 3a and 3b (resident and enumeration), and for "No imputed characteristics" and "1 or more imputed characteristics" (match and enumeration); the variables derived from A.C.E. operations – match code group, housing-unit address match code, and mover status – were retained. Alternative (2) applies the smaller set of cells only in the *P*-Sample, that is, only for unresolved resident and match status; alternative (3) applies it only in the *E*-Sample (enumeration status); and alternative (4) applies it to both samples.

Under alternative (2), the greatest change in the resident probabilities assigned to unresolved cases occurred in the four (original) imputation cells in group 3a, affecting only 96 people with unresolved status. In most other cells for resident status (over 99% of the cases), the probabilities changed very little. A large difference in match probabilities occurred only in the cell "non-mover, nonmatched unit or conflicting household, one or more imputed characteristics," containing 421 unresolved cases. The variable differentiating the number of imputes appears to have had an effect here; if its two "impute" subcells are collapsed, the probability assigned to the "one or more" cell is dominated by the much larger number of resolved people with no imputes, raising the value from 0.567 to 0.684. As is seen in Table 9, the effect on the dual-system estimates is statistically significant for the U.S. total and almost all the breakdowns shown, except for two race-ethnicity groups with sizes under one million people. The relative differences do not appear to be very large, however, ranging from 0.01% to 0.04%. It is not obvious which missing data option produces estimates closer to the unknown true values.

Table 9
Dual-System Estimates Under Alternative Missing Data Procedures

Each cell to the right of the vertical bar contains, in order, estimates of (a) the difference: the alternative estimate minus the A.C.E. estimate, (b) the standard error of that difference, and (c) the relative difference as a percent.

	A.C.E. Estimate (Standard Error)	Estimated Differences Based on Six Alternatives to A.C.E. Missing Data Procedures					
		(1) Noninterview Adjustment With Collapsed Cells	(2) Collapsed Imputation Cells: <i>P</i> -Sample Only	(3) Collapsed Imputation Cells: <i>E</i> -Sample Only	(4) Collapsed Imputation Cells: <i>P</i> and <i>E</i> -Samples	(5) Imputing Probabilities Based on the MES	(6) Imputing Probabilities Based on the MER
U.S. Total	276,848,873 (366,543)	-4,299 (7,423) 0.00%	-55,284 (1,623) -0.02%	-568 (2,581) 0.00%	-55,852 (3,045) -0.02%	-63,632 (5,368) -0.02%	385,969 (24,358) 0.14%
Race-Ethnicity Domains							
Nonhispanic White	194,226,285 (265,893)	-2,467 (6,247) 0.00%	-32,324 (1,055) -0.02%	-1,677 (1,870) 0.00%	-34,000 (2,163) -0.02%	-61,817 (4,534) -0.03%	108,604 (13,026) 0.06%
Nonhispanic Black	34,210,774 (118,415)	-3,495 (1,290) -0.01%	-11,136 (753) -0.03%	-119 (1,328) 0.00%	-11,255 (1,528) -0.03%	-1,303 (1,417) 0.00%	124,710 (11,343) 0.36%
Hispanic	35,552,109 (138,870)	725 (3,016) 0.00%	-8,132 (857) -0.02%	1,432 (973) 0.00%	-6,700 (1,297) -0.02%	196 (1,577) 0.00%	124,937 (10,657) 0.35%
Native Hawaiian or Pacific Islander	618,698 (17,873)	-98 (81) -0.02%	-73 (72) -0.01%	88 (43) 0.01%	15 (85) 0.00%	-107 (74) -0.02%	1,330 (616) 0.22%
Nonhispanic Asian	10,056,009 (64,372)	709 (571) 0.01%	-3,175 (356) -0.03%	-257 (439) 0.00%	-3,431 (567) -0.03%	-414 (576) 0.00%	19,556 (3,704) 0.19%
American Indian on Reservation	567,053 (7,235)	-245 (300) -0.04%	-59 (49) -0.01%	61 (17) 0.01%	2 (52) 0.00%	-38 (73) -0.01%	1,402 (250) 0.25%
American Indian <i>not</i> on Reservation	1,617,944 (22,032)	572 (661) 0.04%	-386 (68) -0.02%	-96 (174) -0.01%	-482 (186) -0.03%	-148 (144) -0.01%	5,430 (1,446) 0.34%
Tenure							
Owner	188,764,543 (260,408)	-2,237 (3,805) 0.00%	-34,503 (1,205) -0.02%	933 (1,971) 0.00%	-33,570 (2,317) -0.02%	-7,816 (1,942) 0.00%	125,058 (10,063) 0.07%
Non-Owner	88,084,330 (226,108)	-2,063 (6,057) 0.00%	-20,782 (1,121) -0.02%	-1,501 (1,607) 0.00%	-22,282 (1,935) -0.03%	-55,816 (5,071) -0.06%	260,911 (21,684) 0.30%
Age Group							
0 - 17	73,076,071 (137,126)	2,924 (2,624) 0.00%	-21,872 (625) -0.03%	-3,315 (1,324) 0.00%	-25,186 (1,474) -0.03%	-8,559 (2,047) -0.01%	107,308 (9,785) 0.15%
18 - 49	129,785,393 (208,070)	-2,721 (4,714) 0.00%	-23,304 (1,143) -0.02%	3,247 (1,565) 0.00%	-20,057 (1,930) -0.02%	-44,534 (3,777) -0.03%	244,070 (16,245) 0.19%
50 and Over	73,987,409 (111,125)	-4,502 (2,766) -0.01%	-10,108 (563) -0.01%	-500 (670) 0.00%	-10,608 (877) -0.01%	-10,538 (1,421) -0.01%	34,591 (4,561) 0.05%

Under alternative (3), the enumeration probabilities were re-computed using only the match code groups as imputation cells. Noticeable changes were detected in the probabilities in the (original) cells for match code group 3a. In the dual-system estimates, the only significant differences were found in two of the three age categories and some of the small race-ethnicity domains. Except for the latter domains, all the percent differences were under 0.01%. As alternative (4) uses the re-computed probabilities from the *P* and *E*-Samples, the resulting estimates here were dominated by the *P*-Sample results and thus were similar to those under alternative (2).

The final two alternative procedures employed the same set of imputation cells as those used in the A.C.E., but assigned to unresolved cases in both the *P* and *E*-Samples potentially improved probabilities, as determined from one of two evaluations conducted by the Census Bureau following the A.C.E. Alternative (5) secured its probabilities from the Matching Error Study (MES), while alternative (6) based them on the Measurement Error Reinterview (MER). Each study took place in a set of evaluation clusters, a roughly one-in-five subsample of the A.C.E. sample block clusters. Information on the MES and MER sample designs can be found in Petroni (2001) and Killion (2000).

The primary purpose of the MES was to evaluate the A.C.E. person matching operation. The evaluation clusters were rematched by expert matchers, and appropriate changes were made to final match codes and person status. No additional data were collected for the MES. Imputation cell probabilities based on MES data were generally similar to those assigned in the A.C.E. One exception, for resident status, was in the cell for match code group 4, Nonhispanic white, non-owner. Here, the MES probability, 0.712, was much lower than the A.C.E. value of 0.911. This resulted from one cluster in the cell that had 24 persons with large weights geocoded incorrectly, as detected in the MES. The MES enumeration probability for match code group 11, "1 or more imputed characteristics," 0.176, was a bit higher than that for the A.C.E., 0.088. Most other probabilities for resident, match, and enumeration status were close (within 0.03) between the A.C.E. and MES; all others were within 0.07.

In contrast, the MER was designed to evaluate the *data collection error* arising from the A.C.E. matching process. People in the MER were reinterviewed about nine months after Census Day to collect information analogous to that collected in the A.C.E. follow-up operation, but in greater detail. Based on the MER, resident probabilities tended to be substantially higher for the cells in match code group 8, but to be lower for the cells in groups 3, 4, and 5 (denoting nonmatches). The reductions tended to be larger in cells where group 8 took more cases away from groups 3, 4, and 5. One might note that the MER cells in subgroup 3a were

fairly small. The "Nonhispanic white, non-owner" cell had only 34 unweighted resolved persons, while the other three cells in group 3a ranged from 125 to 140 unweighted resolved persons. The MER probabilities for enumeration status exhibited similar behavior, with probabilities in groups 11 and 12 raised, and those in the nonmatch groups (3 through 7) lowered. Match probabilities were similar between A.C.E. and MER, mostly differing by 0.01 to 0.05.

Before looking at the dual-system estimates under alternatives (5) (MES probabilities) and (6) (MER probabilities), one should note that, *for the comparison in Table 9*, only the probabilities assigned to unresolved cases were changed based on data collected through the MES or MER. Although the evaluated status of some people may have changed (for example, from nonmatch to match, or confirmed resident to unresolved resident) based on the evaluations, their status was not changed when computing these estimates, as the goal of this exercise was only to explore different methods or information *as they affect the missing data procedures component* of the dual-system estimates.

Under alternative (5), based on MES data and probabilities, the estimates decreased in almost all population domains in Table 9, although never more than 0.1%. Yet this decrease can be attributed almost exclusively to the domain Nonhispanic White. With alternative (6) based on MER data and probabilities, there were significant increases in the estimates of every domain. The relative differences under alternative (6) are larger in magnitude than for earlier alternatives, but all have an absolute magnitude of less than 0.4%. There are several relative differences greater than 0.3% in absolute value: for Nonhispanic Black, Hispanic, and American Indian not on Reservation.

6. OBSERVATIONS

The observations given here pertain to the third type of missing data, assigning probabilities to unresolved people in the A.C.E. It is important to note that the A.C.E. procedures were specified well before the conduct of the census and the A.C.E. The early deadlines were due to (1) the very tight schedule coordinating many separate but interrelated activities, and (2) the need for a process open to the scrutiny of policy makers as well as statistical experts. Although one can learn much about the missing data and the relevant correlation structures by examining the responses as they are collected, making decisions after seeing the data might have been construed as manipulating the results of an operation that had serious political implications.

In this light, one can look back and realize various ways to improve the process – too late to change the procedures.

This does not imply that we did not react to information made available unexpectedly during the processing of the data. We knew that the post-match follow-up operation would help resolve some cases, especially those whose true residence on Census Day was uncertain. Much other information was collected in these interviews, but we did not anticipate seeing the details. However, due to an intensive keying of the follow-up interview forms at the Bureau's processing center, some additional information was made available during the missing data operation. At that time, we added several match code groups not originally in the plan: group 8 for resident status; 11 and 12 for enumeration status. Separating the people in these groups allowed us to assign probabilities that were quite different – and, we believe, more accurate – from what they would have received.

Different models, imputation cells, or data could have been used to assign probabilities for unresolved cases. The values determined through logistic regression were quite similar on average, and may or may not have had an effect on the resulting population estimates. In section 5 it was shown that ignoring some of the demographic variables would have made a difference in the match rate, but probably not in the rate of correct enumeration. Basing the probabilities on data collected in the Matching Error Study or the Measurement Error Reinterview (not yet available during the A.C.E.) could have made a larger difference still. But it is unclear which one might have made an improvement; using MES data would have lowered the population estimates, while using MER data would have increased them.

Weighing the various results, one is constantly reminded that, when assigning probabilities to people with unresolved status, match code group was the most important variable. It worked well for resident and enumeration status, but could not be effectively used for match status. The problem there – perhaps the biggest hole in our procedures – is once again that almost all of the unresolved matches, and over half of the unresolved residents, were people with insufficient information for matching. Little information was collected on these cases, and almost all of them were not sent through the matching process or follow-up. Further, almost none of these people were included in any post-A.C.E. evaluations. In future tests a concerted effort should be made to obtain real information about the status of such people.

ACKNOWLEDGEMENTS

The authors thank Eric Schindler and Doug Olson for computing dual-system estimates and their standard errors under alternative procedures; Tom Belin, UCLA, for making available imputation probabilities under logistic

regression models; and Mary Frances Zelenak and Ha Nguyen for compiling summaries of the extent of missing data in the A.C.E. This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U. S. Census Bureau.

REFERENCES

- ANDERSON, M.J., and FIENBERG, S.E. (1999). *Who Counts? The Politics of Census-Taking in Contemporary America*. New York: The Russell Sage Foundation.
- BELIN, T. (2001). Evaluation of unresolved enumeration status in 2000 Census Accuracy and Coverage Evaluation program. Unpublished report, prepared by Datametrics, Inc., for the U.S. Census Bureau.
- BELIN, T., DIFFENDAL, G., MACK, S., RUBIN, D., SCHAFER, J. and ZASLAVSKY A. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation. *Journal of the American Statistical Association*. 88, 1149-1166.
- CANTWELL, P.J. (2001). Accuracy and Coverage Evaluation Survey: Specifications for the missing data procedures. *DSSD Census 2000 Procedures and Operations Memorandum Series*. Q-62.
- CANTWELL, P.J., MCGRATH, D., NGUYEN, N. and ZELENAK, M.F. (2001). Accuracy and Coverage Evaluation: missing data results. *DSSD Census 2000 Procedures and Operations Memorandum Series*. B-7*.
- CHILDERS, D. (2000). The Design of the Census 2000 Accuracy and Coverage Evaluation. *DSSD Census 2000 Procedures and Operations Memorandum Series*, Chapter S-DT-1.
- FENSTERMAKER, D. (2000). The Accuracy And Coverage Evaluation: sample design summary. *DSSD Census 2000 Procedures and Operations Memorandum Series*. R-33.
- HAINES, D. (2003). A.C.E. Revision II results: changes in estimated net undercount. *DSSD A.C.E. Revision II Memorandum Series*. PP-58
- HOGAN, H. (1993). The Post-Enumeration Survey: Operations and results. *Journal of American Statistical Association*. 88, 1047-1060.
- HOGAN, H. (2003). The Accuracy and Coverage Evaluation: Theory and design. *Survey Methodology*. 29, 129-138.
- IKEDA, M., KEARNEY, A. and PETRONI, R. (1998). Missing data procedures in the Census 2000 Dress Rehearsal Integrated Coverage Measurement sample. *Proceedings of the Survey Research Methods Section*, American Statistical Association. 617-622.
- KEARNEY, A., and IKEDA, M. (1999). Handling of missing data in the Census 2000 Dress Rehearsal Integrated coverage measurement sample. *Proceedings of the Survey Research Section*, American Statistical Association. 468-473.

- KEATHLEY, D., KEARNEY, A. and BELL, W. (2001). ESCAP II, Analysis of missing data alternatives for the Accuracy and Coverage Evaluation. Executive Steering Committee for A.C.E. Policy II (ESCAP II) Report 12.
- KILLION, R.A. (2000). Measurement Error Reinterview Sample Selection. *Planning, Research, and Evaluation Division TXE/2010 Memorandum Series*. CM-MER-S-01.
- KOSTANICH, D. (2001). Accuracy and Coverage Evaluation Survey: computer specifications for Person Dual System Estimation (U.S.) - Re-issue of Q-29. *DSSD Census 2000 Procedures and Operations Memorandum Series*. Q-37.
- NAVARRO, A., and OLSON, D. (2001). Accuracy and Coverage Evaluation: effect of targeted extended search. *DSSD Census 2000 Procedures and Operations Memorandum Series*. B-18*.
- PETERSON, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station*. 6. 1-48.
- PETRONI, R. (1997). Effect of using the 1996 ICM characteristic imputation and probability modeling methodology on the 1995 ICM P and E-sample data. *DSSD Census 2000 Dress Rehearsal Memorandum Series*. A-20.
- PETRONI, R. (1998a). Effect of different methods for calculating match and residence probabilities for the 1995 P-sample data. *DSSD Census 2000 Dress Rehearsal Memorandum Series*. A-23.
- PETRONI, R. (1998b). Effect of different methods for calculating correct enumeration probabilities for the 1995 E-sample data. *DSSD Census 2000 Dress Rehearsal Memorandum Series*. A-28.
- PETRONI, R. (1998c). Effect of using simple ratio methods to calculate P-sample residence probabilities and E-sample correct enumeration probabilities for the 1995 data. *DSSD Census 2000 Dress Rehearsal Memorandum Series*. A-30.
- PETRONI, R. (2001). EFU Sample Design, Stratification, Selection, and Weighting. Planning, Research, and Evaluation Division TXE/2010 Memorandum Series. CM-GES-S-02-R2.
- RUBIN, D.B. (1976). Inference and Missing Data. *Biometrika*. 63, 581-592.
- SEKAR, C.C. and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*. 44, 101-115.

Coverage Error in Population Censuses: The Case of Turkey

H. ÖZTAŞ AYHAN and SÜHENDAN EKNİ¹

ABSTRACT

Coverage errors and other coverage issues related to the population censuses are examined in the light of the recent literature. Especially, when the actual population census count of persons are matched with their corresponding post enumeration survey counts, the aggregated results in a dual record system setting can provide some coverage error statistics. In this paper, the coverage error issues are evaluated and alternative solutions are discussed in the light of the results from the latest Population Census of Turkey. By using the Census and post enumeration survey data, regional comparison of census coverage was also made and has shown greater variability among regions. Some methodological remarks are also made on the possible improvements on the current enumeration procedures.

KEY WORDS: Census coverage error; Coverage error measures; Coverage error estimation; Dual record system estimate; Population census; Post enumeration survey.

1. INTRODUCTION

Coverage has been an important issue in censuses as well as in sample surveys. The difference between the census count and the target population count is the coverage error. When the census count is less than the target population count, it creates an undercount as is common in many countries.

Several techniques are available to understand the problem of coverage errors in censuses. Dual record system (DRS) estimator (Chandra Sekar and Deming 1949) was also extended by many researchers (Ayhan 2000; Casady, Nathan and Sirken 1985; Hogan 1990, 1993a and 1993b; Isaki 1992; Marks, Seltzer and Krotki 1974).

Dual record system estimates based on the census enumeration and a post enumeration survey (PES) are used by the U.S. Census Bureau to measure census coverage error (Hogan 1993a and 1993b; Mulry and Spencer 1988, 1990 and 1993). Post enumeration surveys can be used to improve the population estimate (Ayhan and Eknî 1991; Diffendal 1988; Hogan 1990; Hogan and Wolter 1988).

For the United States, the 1980 Census Post Enumeration Program attempted to measure census coverage through direct measurement using sample survey models (Fay, Passel, Robinson and Cowan 1988). Several methods are also proposed for the adjustment of census count for under enumeration (Choi, Steel and Skinner 1988; Cressie 1988 and 1990).

Recently, models for population coverage error have been studied extensively (Isaki 1992; Wolter 1986). A method of overlapping data systems or multiple frame methodology was used to improve the population estimates (Goodman 1949; Hartley 1962 and 1974; Bankier 1986).

This study highlights the methodological problems related to the population census coverage and proposes

remedies to some of the issues covered. In addition, it proposes and discusses alternative estimates for the population census coverage errors. To achieve the above goals, coverage evaluation issues are included in the design of the PES.

In this paper, section 2 discusses methods of census enumeration and section 3 covers post enumeration survey procedures. Methods of coverage error estimation is presented in section 4 of the paper. Estimators of the population total is given in section 5 and comparison of the coverage error statistics are presented in section 6. Important findings are summarized in the conclusion.

2. METHODS OF CENSUS ENUMERATION

Population censuses have many common features in most countries. The method of enumeration can either be based on *de facto* or *de jure* system. In *de jure* system people are enumerated at their normal residence, while *de facto* system enumerates people actually there. *De facto* system is widely used in developing countries, and the *de jure* system is generally used in developed countries. Countries that are using *de facto* system of enumeration seem to have more problems related to coverage, than the countries which are using *de jure* system of enumeration. These problems stem mainly from their existing imperfect frames for their target population.

De facto based population censuses are generally conducted on a single day, as a complete count, to determine the total population within the country on the day of enumeration. The citizens of the country who are living outside the country were excluded from the census, whereas alien population who are present within the country were included in the census.

¹ H. Öztaş Ayhan, Department of Statistics, Middle East Technical University, 06531 Ankara, Turkey; Sühendan Eknî, State Institute of Statistics, 06100 Ankara, Turkey.

Design of the Census Operations

The Population Census of Turkey was conducted, on the basis of *de facto* system of enumeration, by the State Institute of Statistic (SIS) to determine the quantitative, social and economic characteristics of the population.

For the Census enumeration, the *list of buildings* are created by the local authorities and send to the local Census Committees. Based on the list of buildings (Forms 1 or 2), *enumeration districts* (EDs) *list of buildings* (Form C) are formed (see Appendix 1 for details). Due to the lack of timely availability of the complete list of buildings in the SIS Central Office, the number of EDs are estimated by projection techniques for advance fieldwork planning of the Census, as well as PES. EDs are obtained by assigning 100 persons per enumerator in province and district centers, and 200 persons per enumerator in sub-districts and villages, based on average daily workloads. They are then numbered sequentially. In the Census, the listed addresses were taken as the base for identifying the “dwelling units (DUs)”, while the “individual persons” within the household(s) of the dwelling unit is considered as the unit of enumeration.

The workload of each enumerator is taken as an ED, which contains a list of addresses to be covered within a specified close interval. Instructions are given to the enumerator to treat this interval as a compact segment. If an enumerator encountered an address not on the list, it is included in the enumeration by work definition. For vacant and nonexistent units the related information is also recorded. There was no special procedure for dealing with reluctant respondents or in general any non-interviewed units, due to the compulsory nature of response by the related Act. The enumerator's workload is set in such a way that, they will complete all the interviews in a given day. For very special cases, the instruction is given to complete the enumeration of the segment during the extended hours in the same census day.

Additional enumerators were assigned to enumerate the “*special enumeration districts*” such as the places of the mobile populations (travelers, persons on duty, nomadic tribes, etc.) and institutional populations (hospitals, prisons, factories, military establishments, etc).

Institutional population are covered by additional enumerators, who are assigned for these special EDs. The mobile populations travelling by vehicles are stopped and were enumerated as a group when they first appeared within the borders of the provinces. The passengers continues their journey after enumeration and duplicate enumerations are avoided by placing an “*Enumerated*” sign on the vehicle after the census operation, and later their individual identification is also checked by manual and computerized algorithms, against the other records of the relevant settlement.

The Census was conducted on a Sunday, and the enumeration was completed on the same day. On the Census day, a national curfew was declared. The

enumerators visited each household (HH) within the dwelling units listed in their enumeration district building lists and completed the census questionnaire (see Appendix 2 for details). For the *Household Module*, the information is collected from an adult household member for the general household characteristics, while for the *Individual Person's Module* the information is obtained from self respondents on their personal characteristics.

The following type of errors occurred during the different stages of the census operations;

- (1) Omission errors and erroneous inclusions has occurred during the construction of the List of Buildings. However, due to the use of compact segment approach in the enumeration process of the census operation, these errors are mostly eliminated.
- (2) Response errors based on memory recall error, cheating, and inadequate answer for coding has occurred during the census enumeration. These are measured as the response inconsistency in the *Response Reliability Study* (Ayhan and Ekni 1991; SIS 1994) of the Population Census, which was based on the PES.
- (3) Some enumerator errors (failure to probe, inadequate perception of response, and recording errors) are also observed during the census operation. These are also covered by the *Response Reliability Study*.
- (4) Processing errors such as, coder and verifier errors also occur during the data processing and these are eliminated later during the data handling in the office.

3. POST ENUMERATION SURVEY PROCEDURES

The objectives of the PES are to determine coverage error in the population census as well as obtain measures of response reliability of the questions in the census. In this paper, the first objective is discussed for the Population Census of Turkey, and the preliminary findings for both objectives are summarized by Ayhan and Ekni (1991).

3.1 Sample Selection Procedures

The sample design for the PES is initiated 3 months before the Census operation. At this stage, creation of the Population Census EDs was not complete yet.

Stratification and estimation of population EDs. The previous Population Census enumeration district lists of the State Institute of Statistics is used as the base for sampling frame for PES operation. The population of people is first stratified into 5 *geographical-socio economical regions* of Turkey. A second explicit stratification variable is also used, which is based on the 8 *size groups for the place of the settlements*, in a nested structure within regions. Here,

urban-rural boundary corresponds to a population size of 10,000. The number of census enumeration districts were estimated for 40 *design strata* for the Census day by using forward population projection method, which were based on the person counts of the two previous population censuses. For the census enumeration, EDs were created by using Form C within the Central Office. A total of 479 251 EDs are established for the Census. Sampling frame information is given in Table 1.

Table 1
Estimated Number of Population and Sample EDs by
Regions and Urban-Rural Strata

Region	URBAN		RURAL		TOTAL	
	Popu. ED	Samp. ED	Popu. ED	Samp. ED	Popu. ED	Samp. ED
h	$M_h^{(U)}$	$m_h^{(U)}$	$M_h^{(R)}$	$m_h^{(R)}$	M_h	$m_h^{(1)}$
1	125,726	125	40,333	32	166,059	157
2	42,442	42	24,992	20	67,434	62
3	65,466	76	45,925	36	111,391	112
4	15,790	16	30,459	22	46,249	38
5	39,358	40	48,760	34	88,118	74
Total	288,782	299	190,469	144	479,251	443

The expansion factors: $F_h^{(1)} = N_h^{(1)} / n_h^{(1)} \neq M_h / m_h^{(1)} = F_h^{(2)}$

The coverage of the number of dwelling units in the Census and PES were achieved by the following procedures. The number of population EDs for each province was determined and numbered sequentially. Then, the number of population EDs in each population strata was estimated by, dividing the projected strata population (N_h) to the fixed daily workload of enumerators (B_{hi}). Population EDs were estimated for urban areas as $M_{hi} = N_{hi} / B_{hi}$ and for rural areas as $M_{hj} = N_{hj} / B_{hj}$, where the ED sizes are taken as fixed daily workload, $B_{hi} = 100$ persons in urban strata and $B_{hj} = 200$ persons in rural strata. The results of the population projections for each strata by urban-rural aggregation are also obtained. The estimated number of population EDs and expansion factors for regions and urban-rural strata are also computed.

Selection of sample EDs. A stratified multistage sample of localities and blocks are selected systematically for PES from the available *master sampling frame* of the State Institute of Statistics at the Central Office. The blocks of the master sampling frame is periodically updated for the multi purpose selection of other samples on routine basis. The interviewers of the PES is recruited and trained in the Central Office, and then interviewers are send as a team to the local sample settlements for the independent enumeration of the selected PES sample. For the identification purpose, the selected sample blocks are linked to their corresponding Population Census EDs of the settlement in the field by previously given instructions to the PES interviewers.

For the use of Dual Record System estimation, the sample enumeration districts for the PES should be determined independently from the census frame. This is an absolutely crucial assumption of the DRS model, which was emphasised by many researchers during the past 50 years (Ayhan 2000; Chandra *et al.* 1949).

Due to the use of unwanted old ED lists in some areas, the range of the planned workload per ED per enumerator may have changed and consequently the selected sizes of the EDs may be different from the actually enumerated sizes. This will effect the achieved sampling fractions, which will naturally be different from the selected.

A total of $m = 443$ sample enumeration districts are selected in 16 province centers, 23 districts, 16 sub-districts and 43 villages within the 40 strata. For the PES, a sample of 443 EDs are selected from the created ED list by systematic sampling.

The sampling fractions and sample allocation was achieved in the following way. Equal probability selection method was used to select the sample enumeration districts in all strata. Sampling fractions were planned to be $f_h \approx 0.001$ for all strata. However, the sampling fractions are also varied among strata. Technical details of the sampling fractions and the sample allocation are given below. The sampling fractions [$f_h^{(1)}$] and sample allocation [$n_h^{(1)}$] can be achieved as,

$$f_h^{(1)} = n_h^{(1)} / N_h^{(1)} = 1 / F_h^{(1)} \text{ and } f_h^{(2)} = m_h^{(1)} / M_h = 1 / F_h^{(2)}. \quad (1)$$

The total population sizes of urban (U) and rural (R) EDs are,

$$N_h^{(U)} = M_h^{(U)} B_{hi} = \left[\sum_i^I M_{hi} \right] B_{hi} \quad \forall h \text{ \& } i \text{ and} \quad (2)$$

$$N_h^{(R)} = M_h^{(R)} B_{hj} = \left[\sum_j^J M_{hj} \right] B_{hj} \quad \forall h \text{ \& } j \quad (3)$$

where the components are defined earlier.

Then the population size of each stratum was determined as

$$N_h^{(1)} = \left[N_h^{(U)} + N_h^{(R)} \right]. \quad (4)$$

Similarly, the corresponding sample sizes of each stratum are

$$n_h^{(1)} = \left[n_h^{(U)} + n_h^{(R)} \right] \quad (5)$$

$$\text{where } n_h^{(U)} = m_h^{(U)} B_{hi} \text{ and } n_h^{(R)} = m_h^{(R)} B_{hj}. \quad (6)$$

3.2 Design of the PES Operations

The fieldwork operation for PES was identical to the census, where the details are given in section 2.2. For operational purposes, each ED was defined as a close

interval of dwelling unit numbers within the streets. In terms of special enumeration districts (*i.e.*, institution) the total number of enumeration districts are checked with prior information which was obtained at province level.

Due to previously given instructions to the enumerators, PES starts in the sample enumeration districts an hour after the starting time of the census enumeration on the same day. PES enumerators visit the same households in the same (ascending) order as the census enumerators, so that PES enumerators did not visit the same household before the census enumerators. Results of the PES was used as a basis for evaluation, after matching the individual cases with the census records for the corresponding EDs.

4. METHODS OF COVERAGE ERROR ESTIMATION

This section addressed coverage error estimation, by stating data matching procedures and dual system estimation procedures and related findings. The evaluation and estimation of population coverage is obtained using the list of EDs from two independent sources. In this section, the data matching procedures, dual record system estimators, alternative population total estimators are proposed and the estimates are evaluated. A comparison of the computed coverage error statistics are also presented here.

4.1 Data Matching Procedures

Several models (Deming and Glasser 1959; Nathan 1967 and Tepping 1968) have been proposed for determining the optimum matching procedures. These are based on establishing procedures that minimise the “*estimated net matching error*” subject to given costs and other constraints (Marks *et al.* 1974). These models provided valuable concepts to the theory and practice of matching, but none of are completely satisfactory for all purposes.

The work of Tepping (1968), extended by Srinivasan and Muthiah (1968), required a minimum set of characteristics for the “*exact agreement*” in matching. Also, Ayhan and Eknî (1991) and SIS (1994) have used similar methods based on the following specifications;

- (1) *Matching of the population of the EDs.* The total population of the ED was taken as the sum of the household population within the total DUs of the ED.
- (2) *Matching of the households within the EDs.* Several sets of information (address of the dwelling unit, names of household head and number of persons in the household) was considered for matching of households.
- (3) *Matching of the individual persons within the matched households.* A total of 4 Census / PES variables (names, age, sex and education level) were

all used for exact agreement in matching of individuals.

- (4) *Matching of non-matched individuals of the households.* This was achieved by matching with the other individuals in the neighboring households (from the other data source) by searching. The same criteria was used for exact agreement in matching of individuals.

The preliminary work of matching operation is done clerically, while matched households and persons are evaluated by automation. For the matching procedure of persons the frequencies $n(r, c)$ are shown in Table 2.

Table 2
The Layout of the Matching Procedure

		DATA SOURCE 2: (PES)		
Matching Procedure		In	Not in	Total
DATA	In	$n(1, 1)$	$n(1, 2)$	$n(1, *)$
SOURCE 1:				
(CENSUS)	Not in	$n(2, 1)$	$\hat{n}(2, 2)$	$n(2, *)$
Total		$n(*, 1)$	$n(*, 2)$	\hat{n}

On the basis of the above specifications, the households are matched at the first stage, and within the matched households the persons are matched at the second stage. The results are presented in the following tables by regions. Enumeration districts are located in sample settlements within 19 provinces in 5 regions of the sample design. Out of 443 selected EDs, 437 were matched with their corresponding population census counterparts and other 6 EDs could not be matched due to differences in independently given instructions for their creation by the local offices. The information on the regional breakdown of the 6 non-matched EDs are provided in Table 3, while the information on the urban-rural breakdown was not obtained.

The matching procedure of households can be illustrated by $k(r, c)$ in the same way as presented for persons in Table 3. In the stratified case, the number of households in each strata can also be denoted by $k_h(r, c)$. The total number of households in the Census which are not matched with PES households can be estimated for each strata as

$$k_h(1, 2) = [k_h(1, *) - k_h(1, 1)] \quad (7)$$

and the total number of households in the PES which are not matched the Census households can also be estimated for each strata as

$$k_h(2, 1) = [k_h(*, 1) - k_h(1, 1)]. \quad (8)$$

Information on matched and non-matched households are given in Table 3.

Table 3

Matched and Non-matched Households in the Post Enumeration Survey and Census Enumeration Districts by Regions

Regions h	Selected no. of EDs $m_h^{(1)}$	Enumerated no. of EDs $m_h^{(2)}$	Matched HHs $k_h(1, 1)$	Non-matched households	
				Census $k_h(1, 2)$	PES $k_h(2, 1)$
1	157	154	3,320	168	144
2	62	62	1,262	27	30
3	112	112	2,636	262	259
4	38	38	645	204	80
5	74	71	995	170	175
Total	443	437	8,858	831	688

In these 437 EDs, a total of 8858 households were matched. In the Census 831 (9.38 %) and in PES 688 (7.77%) households could not be matched. The Census based household match rate was 90.62 %, while PES based match rate was 92.23 %, which is presented in Table 3.

Coverage rates for the Census and PES households are given by regions in Table 4. Comparison of Census and post enumeration results for the coverage rate of households (C_h) were higher for the Census in most regions (except Regions 2 and 5) and the total. Here all the coverage rates were greater than it was expected. In terms of persons within the covered households, the coverage rates (C_h^*) were higher for the Census for all regions and for the total. Total of matched persons were $n(1, 1) = 41,020$ in the Census and PES.

Differences in the coverage of EDs in the Turkish Census and PES comes due to the following reasons;

- (1) Additional Forms of C and D are established by the Census Committee of the provinces through list of buildings (Forms 1 and 2). List of buildings are created by the local authorities and they are not reliable enough for some settlements.
- (2) Numbering of EDs are also done at the local level, they are also effected by the insufficient numbering operation.

- (3) Forms C and D may or may not contain 100 persons in urban and 200 persons in rural areas due to outdated listings.
- (4) Due to different starting points by the Census and PES enumerators, the number of dwelling units covered were different.
- (5) Application of the PES questionnaire was started at least 2 hours after the actual Census operation within the selected EDs. Coverage differences may be due to the mobility of the members of census completed households within the same ED.
- (6) During the one day enumeration period, some of the planned Census and PES questionnaires could not be completed, resulting inconsistency during matching. Of course, this is a source of undercount, which happened rarely during the field enumeration.
- (7) Because of the de facto enumeration base, the local visitors (from other dwellings of the apartment) for either data source were subject to change.
- (8) Again, due to de facto enumeration, there will be counting errors for the mobile population for the Census. The PES only planned to cover the household population.
- (9) The PES was not planned to cover the special EDs and mobile populations (*i.e.*, travelers, persons on duty, *etc.*). By definition, international and domestic travelers are permitted to continue their travel after being counted, if their journey had started before the official census starting time. During this research, the mobile population was excluded from the analysis.
- (10) Nomadic tribes (Special enumeration techniques are required for the census of nomadic tribes. *De jure* rather than *de facto* enumeration base, as well as mobile interviewers may be recommended for the enumeration of nomadic tribes in place of interviewers who are stationary.) will not be covered in the PES due to non-listings.

Table 4

Number of Households and Persons in the Census and Post Enumeration Survey by Regions

Regions h	Number of Household			Number of Persons in Households			
	Census $k_h(1, *)$	PES $k_h(*, 1)$	Coverage C_h	Census $n_h(1, *)$	PES $n_h(*, 1)$	Matched $n_h(1, 1)$	Coverage C_h^*
1	3,488	3,464	1.0069	14,035	13,926	13,393	1.0078
2	1,289	1,292	0.9977	6,587	6,582	6,400	1.0008
3	2,898	2,895	1.0010	13,058	12,984	11,644	1.0057
4	849	725	1.1710	4,233	3,580	3,134	1.1824
5	1,165	1,170	0.9957	7,898	7,888	6,449	1.0013
Total	9,689	9,546	1.0150	45,811	44,960	41,020	1.0189

Coverage rates: $C_h = k_h(1, *) / k_h(*, 1)$ and $C_h^* = n(1, *) / n(*, 1)$

- (11) Both Census and PES EDs are enumerated with the same instruction for the previously defined close interval. However, due to the use of different quality of the frames (updated 1990 or outdated 1985 or even outdated 1980) the amount of workload of each interviewer was varying. Consequently, the amount of coverage in each ED may be different from both sources.

4.2 Dual Record System Estimation

Dual record system is used as a method for determining the estimated number of households and persons through a matching procedure. The results are used to estimate the total number of persons in each region and the total population. The model assumes independence of data collection from two sources, where the Census and the PES are used. In theory, all cells $[n(r, c)]$ are observable except for $n(2, 2)$ and any of the totals that include $n(2, 2)$. Chandra *et al.* (1949) assumes that, there is no correlation bias with the estimate for cell $n(2, 2)$. For practical purposes, this paper also considers this assumption as valid. On the other hand, further discussion on the validity of such assumption is recently reported by Ayhan (2000).

The methodology and the estimation procedures are presented below. Estimation of the number of persons not in the Census or in PES

$$n(2, 2) = [n(1, 2) n(2, 1)] / n(1, 1). \quad (9)$$

Total number of persons is estimated as

$$n = n(1, 1) + n(1, 2) + n(2, 1) + n(2, 2) \quad (10)$$

or alternatively,

$$n = [n(*, 1) n(1, *)] / n(1, 1). \quad (11)$$

Table 2 earlier illustrated the matching procedure used for the dual record system method. The computational procedure presented here was repeated for each region separately. The estimates are given in Table 5. For each strata, n_h is computed as n previously.

Table 5
Matched and Non-matched Number of Persons in the Census and Post Enumeration Survey by Regions

Regions	Matched	Census non-match	PES non-match	Estimated omissions in both sources	Dual record system estimate
h	$n_h(1, 1)$	$n_h(1, 2)$	$n_h(2, 1)$	$n_h(2, 2)$	$n_h^{(D)}$
1	13,393	642	533	26	14,594
2	6,400	187	182	5	6,774
3	11,644	1,414	1,340	163	14,561
4	3,134	1,099	446	156	4,835
5	6,449	1,449	1,439	323	9,660
Total	41,020	4,791	3,940	673	50,424

4.3 Total Population versus Household Population

The total population was considered as the target population for the population projections, which was used to estimate the total number of EDs in the population. On the other hand, PES sample design only considered the household population as the target population. Because the PES design was based on the selected sample dwelling units only, which excluded the special enumeration districts (the institutional population).

As stated earlier, the PES sample design was taken as the base for the comparison of two different enumeration systems during the matching procedures. This naturally led us to consider the household population as the target population for the appropriate estimation of the population total by the proposed estimators. In order to achieve this, the institutional population was computed later, from the 1990 Census data, for regions and population size groups. The institutional population of regions are presented (by aggregating over the size groups) in Table 6.

Table 6
Determination of Household Population and Sample Sizes by Regions

Regions	Projected population size	Institutional population estimate	Household population size	Household survey sample size	Expansion factors	
h	$N_h^{(1)}$	$N_h^{(2)}$	$N_h^{(3)}$	$n_h^{(1)}$	$F_h^{(1)}$	$F_h^{(3)}$
1	20,639,200	367,184	20,272,016	18,900	1092.02	1072.59
2	9,242,600	89,934	9,152,666	8,200	1127.15	1116.18
3	15,731,600	176,031	15,555,569	14,800	1062.95	1051.05
4	7,670,800	55,104	7,615,696	6,000	1278.47	1269.28
5	13,687,800	249,309	13,438,491	10,800	1267.39	1244.30
Total	66,972,000	937,562	66,034,438	58,700	1140.92	1124.95

$$N_h^{(3)} = N_h^{(1)} - N_h^{(2)} \text{ here } F_h^{(1)} = N_h^{(1)} / n_h^{(1)} \text{ and } F_h^{(3)} = N_h^{(3)} / n_h^{(1)}$$

For the further use of the information on the institutional population, it was also assumed that, there were no coverage errors associated in measuring the institutional population during the 1990 Census enumeration. The household population of each region, are then computed by subtraction.

There were several reasons for removing the institutional population from the total population;

- (1) The PES sample design only reflected the household population.
- (2) The correct selection probabilities for the ideal coverage (representation) of each sample strata can only be based on the household population, not on the total population.
- (3) The proposed coverage error estimates should only be based on the household population.
- (4) The proposed estimators for the population total should also be based on the household population, where the PES results are household based.
- (5) It will be wrong and misleading to make comparison of coverage error statistics, when the base populations are different.
- (6) The census undercount is artificially inflated if the wrong population (namely, the total population) is taken as the target population.

4.4 Coverage Error Measures

Many coverage error statistics are proposed in the literature. Some of these error statistics are based on simple ratios or proportions, and others are based on more complex adjustment procedures. To simplify the solution to this problem, the following coverage error measures are proposed for the regional and total population. These are census coverage rate, census discrepancy rate and the amount of census discrepancy. The following coverage error measures are proposed which are based on the household population.

Census Coverage Rate:

Regional estimators:

$$\lambda_h^{(s)} = N_h^* / \hat{N}_h^{(s)} \quad \forall h \quad h = 1, 2, \dots, H \quad (12)$$

where N_h^* = Census count of the household population [$N_h^* = N_h - N_h^{(2)}$] and $\hat{N}_h^{(s)}$ = Estimate from source (or method) s .

Standard error of regional estimators: Making the following scale transformation $\lambda_h^{(s)}(0.5) = \tilde{\lambda}_h^{(s)}$ which is taken as a proportion, realizing that within each strata $\tilde{\lambda}_h^{(s)} + (1 - \tilde{\lambda}_h^{(s)}) = 1$, the standard error estimators of the census coverage rates of each region is computed as

$$se[\tilde{\lambda}_h^{(s)}] = \left[\frac{\tilde{\lambda}_h^{(s)} - (1 - \tilde{\lambda}_h^{(s)})}{n_h^{(D)} - 1} \right]^{1/2} \quad (13)$$

$$\text{Total population estimator: } \lambda = N^* / \hat{N}^{(s)} \quad (14)$$

Census discrepancy rate:

$$\text{Regional estimators: } \varphi_h^{(s)} = 1 - \lambda_h^{(s)} = \left[\delta_h^{(s)} / \hat{N}_h^{(s)} \right] \quad (15)$$

$$\text{Total population estimator: } \varphi = 1 - [N^* / \hat{N}^{(s)}] = 1 - \lambda \quad (16)$$

Census discrepancy:

$$\text{Regional estimators: } \delta_h^{(s)} = \hat{N}_h^{(s)} - N_h^* \quad \forall h \quad (17)$$

Due to the limitations of the one day enumeration by the *de facto* system, other additional local coverage measures could not be considered for this study. Such additional coverage measures for the local areas could provide useful additional information for more complex coverage error models in countries who are employing *de jure* system of enumeration in their census taking.

Even with the limitations of the *de facto* census, one could compute coverage estimates for large domains (such as provinces), where the population would not likely to shift very much between Census and PES interview. This was not possible, due to the limited sample size of PES which did not provide independent provincial estimates to be made within the regions. In addition, the sample sizes might not be large enough to give sufficient precision.

5. ESTIMATORS OF POPULATION TOTAL

The *estimated population total* is taken as the weighted sum of the all regional estimates.

$$\hat{N}^{(s)} = \sum_h^H \hat{N}_h^{(s)} \quad (18)$$

The *standard error estimators* for the total household population of each region is computed as

$$se[\hat{N}_h^{(s)}] = \hat{N}_h^{(s)} \left[\frac{p_h(1 - p_h)}{n_h^{(D)} - 1} \right]^{1/2} \quad (19)$$

while the proportion of each strata is computed as $p_h = n_h^{(D)} / \sum_h^H n_h^{(D)}$.

The determination of the coverage error of a given Census is not an easy task, especially when a perfect list of a target population is not available to compare the results. This is always the case in most countries of the world, except the ones with population registers.

Table 7
Estimates of the Regional and Total Household Population for 1990 by the Expanded Dual Record System
Estimate and Their Standard Errors

h	$\hat{N}_h^{(1)}$	$se[\hat{N}_h^{(1)}]$	$\hat{N}_h^{(2)}$	$se[\hat{N}_h^{(2)}]$	$\hat{N}_h^{(3)}$	$se[\hat{N}_h^{(3)}]$
1	15,936,939	58,967*	15,436,073	57,113	15,653,378	57,917
2	7,635,314	31,305	7,367,741	30,208	7,561,003	31,000
3	15,573,280	57,621	14,571,298	53,914	15,398,933	56,976
4	6,181,402	38,943	5,884,582	37,073	6,136,969	38,663
5	12,242,987	48,972	11,502,934	46,012	12,019,938	48,080
Total	57,569,922	241,794	54,762,628	230,003	56,770,221	238,435

*: Standard error estimates are rounded to the nearest integer.

Comparison of the results of a population census with projection figures also creates some kind of comparison problems, due to the validity of the several assumptions relating to projection models. In order to avoid a single base of comparison, the following *expanded dual record system regional estimators* are proposed for the determination of the census coverage errors.

Estimator 1. $\hat{N}_h^{(1)} = F_h^{(1)} n_h^{(D)}$ (20)

where $F_h^{(1)} = N_h^{(1)} / n_h^{(1)}$ and $n_h^{(D)} = \sum_r \sum_c n_h(r, c)$.

Here $n_h^{(D)}$ refers to the unweighted DRS estimate and $n_h^{(1)}$ corresponds to the selected sample size.

Estimator 2. $\hat{N}_h^{(2)} = F_h^{(2)} n_h^{(D)}$ where $F_h^{(2)} = M_h / m_h^{(1)}$. (21)

Estimator 3. $\hat{N}_h^{(3)} = F_h^{(3)} n_h^{(D)}$ where $F_h^{(3)} = N_h^{(3)} / n_h^{(1)}$. (22)

The dual record system estimators are expected to yield higher estimated counts than a single round survey (*i.e.*, PES), by definition. Therefore, all the proposed estimators for the household population totals are DRS based. DRS estimates of the total household populations are given in Table 7.

Difference between the three proposed estimates, are only based on the type of expansion factors used. When we examine the expansion factors, $F_h^{(1)}$ is based on the projected population sizes over household survey sample sizes. On the other hand, $F_h^{(2)}$ is based on total population EDs over total sample EDs of the original PES design.

Finally, $F_h^{(3)}$ is based on the household population size over household survey sample size. The first two estimators include institutional population components $[N_h^{(2)}]$ in the numerator of their expansion factors $[N_h^{(1)} \text{ or } M_h]$, while only the third estimator uses household population information $[N_h^{(3)}]$ in its expansion factor. It is clear that, the expansion factor for the third estimator is derived from

the ideal selection probabilities $[f_h^{(3)} = n_h^{(1)} / N_h^{(3)} = 1 / F_h^{(3)}]$ for the PES sample, which is based on household information. Therefore, *Estimator 3* can be considered as more representative of the household population.

6. COMPARISON OF COVERAGE ERROR STATISTICS

For the comparison of error statistics, the population counts should be of the same standard base. It will be recommended to use a household population count which matches the corresponding population estimate. The regional and total population counts are given in Table 8. As mentioned earlier, the institutional population counts are determined from the 1990 Census counts.

Table 8
Regional and Total Population Counts for Turkey, 1990

	Census counts	Institutional population counts	Household population counts
h	N_h	$N_h^{(2)}$	N_h^*
1	18,544,967	367,184	18,177,783
2	7,836,940	89,934	7,747,006
3	12,824,347	176,031	12,648,316
4	5,964,565	55,104	5,909,461
5	11,302,216	249,309	11,052,907
Total	56,473,035	937,562	55,535,473

where $N_h^* = N_h - N_h^{(2)}$

For the purpose of population coverage error evaluation, the *census coverage rate* and the amount of *census discrepancy* was used. The computed population coverage error rates are given by regions and the total in Table 9.

Table 9

Estimates of the Census Coverage Rates for Regional and Total Household Population in Turkey 1990 and Their Standard Errors

h	$\lambda_h^{(1)}$	$se[\hat{\lambda}_h^{(1)}]$	$\lambda_h^{(2)}$	$se[\hat{\lambda}_h^{(2)}]$	$\lambda_h^{(3)}$	$se[\hat{\lambda}_h^{(3)}]$
1	1.14061	0.00410	1.17762	0.00407	1.16127	0.00408
2	1.01463	0.00607	1.05148	0.00607	1.02460	0.00607
3	0.81218	0.00407	0.86803	0.00411	0.82138	0.00408
4	0.95601	0.00718	1.00423	0.00719	0.96293	0.00719
5	0.90272	0.00506	0.96088	0.00508	0.91955	0.00507
Total	0.96466	0.00223	1.01411	0.00223	0.97825	0.00223

There is a clear pattern for certain regions, for all estimates. The census coverage rates can also be expressed as the amount of census discrepancy. A similar pattern is expected for the three estimators, since the estimators are highly correlated.

For the total population, estimates based on methods (1) and (3) has resulted in census undercount when compared with the corresponding actual population counts. Due to the computational procedures, *Estimate 3* can be recommended among others because *Estimate 3* is based on the projected household population, where the comparison base is the same as the selection.

There is also a pattern for regional estimates, regardless of the method of estimation. For regions 1 and 2, all estimates indicated census overcount, while census undercount was observed for all other regions by all estimates, except for *Estimate 2* in region 4.

7. CONCLUSIONS

The coverage error study of the population census had provided some useful information in evaluating the methodological issues which is summarised below.

Comparison of the three proposed population total estimates indicates that, the first estimate provided the highest value of the total count, while *Estimate 3* provided more representative result for the total household population.

The evaluation of the census coverage error rates and the amount of census discrepancy had shown that, for the total population, *Estimates 1 and 3* has resulted in census undercount. There is also a distinct pattern for regional estimates, regardless of the method of estimation. There seems to be a census overcount in the first two regions, while census undercount was observed for the other three regions by all estimates (except for *Estimate 2* in Region 4).

For the developing countries, the main problem of census taking is based on the undercount. In Turkey, the overcount issues in census taking only occur in very limited local areas and they are re-evaluated later and removed from the census data before release of the census results.

On the basis of these findings it is clear that, the comparison of several sample based estimates with the

population census count indicated the existence of some methodological problems which are present in the enumeration procedures of the Turkish Population Census. The most important of these issues are the following;

- (1) Improving and updating the list of possible EDs in rapidly growing peripheries of the large cities by the use of area methods.
- (2) Obtaining a perfect list of all dwelling units within the EDs. This can be better achieved through a continuous screening operation by the local authorities, where they are responsible for this by law. Alternatively, a Census of Housing can be taken just before the population census by the SIS which will also provide a useful frame for the population census enumeration.
- (3) There are many laws in the country which refers to the latest "population counts". This suggests that, major changes might be necessary on legal issues as well as in enumeration techniques.
- (4) Enumeration of the mobile populations also requires special attention, methods and qualified personnel.

One would like to hope that, measuring the characteristics of the population through the Censuses may be considered important, by the responsible officials in time and the necessary developments will take place along these directions.

ACKNOWLEDGEMENTS

We would like to thank Professors Orhan Güvenen, Yalçın Tuncer, Vijay K. Verma, Moti Lal Tiku, M. Qamar Islam and Mr. Ömer Gücelioglu for their valuable comments. The contributions of Ms. Hasibe Dedes and Ms. Canan Bakici are also gratefully acknowledged. Finally, the comments and suggestions of the Editor, Associate Editor and the referee's of the Journal is very much appreciated. The views expressed are attributed to the authors and do not necessarily reflect those of the State Institute of Statistics, Turkey.

APPENDICES: TOOLS OF ENUMERATION

The following listing forms and questionnaires are used before and during the Census and PES operations.

APPENDIX 1. LISTING FORMS USED

Form 1: List of Buildings (for localities with municipal organization).

This list is created by the local municipality personnel and later produced in triplicate. Used for sequential numbering of DUs in urban areas.

Form 2: List of Buildings (for localities without municipal organization).

This list is created by the village head person and later produced in triplicate. Used for sequential numbering of dwelling units in rural areas.

Form C: Enumeration District List of Buildings.

This list is based on Forms 1 or 2. The EDs are formed on the basis of this list in urban and rural areas, separately.

Form D: Census Control List.

This is an update of Form C which was completed by the enumerator after the census field operation and returned to the Local Census Committees with the completed census questionnaires. This form and the completed census questionnaires are forwarded to the SIS after the census field operation.

APPENDIX 2. QUESTIONNAIRES USED

Form A: Population Census Questionnaire.

The population census questionnaire consisted of four main parts. The information is collected through a personal interview by a paper and pencil approach.

Part 1. *Address details.*

Part 2. *Type of place of the residence.*

Part 3. *Household module* [contains 7 precoded household questions].

Information is collected to identify the household head, presence of head, total number of persons in HH, number of guests, number of HH members absent, ownership of present DU, and ownership of any other DU.

Part 4. *Individual person's module* [contains 26 precoded individual questions].

For each person present, information is obtained on sex, age, relation to HH head, place of birth, citizenship, permanent residence, educational background, marital status, fertility information, employment status, and main occupation.

Form B: Post Enumeration Survey Questionnaire.

PES questionnaires are generally based on a subset of questions of the main study. However, for this study it was decided by the Census Advisory Committee to use the complete census questionnaire for the PES. The questionnaire for PES is completed in the same way as the Census.

REFERENCES

- AYHAN, H.Ö., and EKNİ, S. (1991). Coverage and response errors in 1990 Turkish Census of Population. *Bulletin of the International Statistical Institute*. 54, 45-46.
- AYHAN, H.Ö. (2000). Estimators of vital events in dual-record systems. *Journal of Applied Statistics*. 27, 157-169.
- BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*. 81, 1074-79.
- CASADY, R.J., NATHAN, G. and SIRKEN, M.G. (1985). Alternative dual system network estimators. *International Statistical Review*. 53, 183-197.
- CHANDRA SEKAR, C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extend of registration. *Journal of the American Statistical Association*. 44, 101-115.
- CHOI, C.Y. STEEL, D.G. and SKINNER, T.J. (1988). Adjusting the 1986 Australian census count for under enumeration. *Survey Methodology*. 14, 173-189.
- CRESSIE, N. (1988). When are census counts improved by adjustment? *Survey Methodology*. 14, 191-208.
- CRESSIE, N. (1990). Weighted smoothing of estimated undercount. U.S. Bureau of the Census, *1990 Annual Research Conference Proceedings*. 301-325.
- DEMING, W.E., and GLASSER, G.J. (1959). On the problem of matching lists by samples. *Journal of the American Statistical Association*. 54, 403-415.
- DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in Los Angeles County. *Survey Methodology*. 14, 71-86.
- FAY, R.E. PASSEL, J.S. ROBINSON, J.G. and COWAN, C.D. (1988). The Coverage of Population in the 1980 Census. *Evaluation and Research Reports*, PHC 80-E4, U.S. Bureau of the Census. 123.
- GOODMAN, L.A. (1949). On the estimation of the number of classes in a population. *Annals of Mathematical Statistics*. 20, 572-579.
- HARTLEY, H.O. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section*, American Statistical Association. 203-206
- HARTLEY, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, Series C. 36, 99-118.
- HOGAN, H. (1990). The 1990 Post Enumeration Survey: An Overview. *U. S. Bureau of the Census Paper*, Washington DC. 6.
- HOGAN, H. (1993a). The 1990 post enumeration survey: operations and results. *Journal of the American Statistical Association*. 88, 1047-1060.
- HOGAN, H. (1993b). Planning for census correction: the 1990 United States experience. Invited Paper, 49th Session of the International Statistical Institute, Florence, Italy. *International Association of Survey Statisticians Booklet*. 133-150.
- HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a post enumeration survey. *Survey Methodology*. 14, 99-116.
- ISAKI, C.T. (1992). Model bias effects in small area coverage error estimation. *Communication in Statistics Serie A*. 21, 1213-1231.
- MARKS, E.S., SELTZER, W. and KROTKI, K.J. (1974). *Population Growth Estimation: A Handbook of Vital Statistics Measurement*. New York: The Population Council.

- MULRY, M.H., and SPENCER, B.D. (1988). Total error in the dual system estimator: the 1986 census of central Los Angeles county. *Survey Methodology*. 14, 241-263.
- MULRY, M.H., and SPENCER, B.D. (1990). Total error in post enumeration survey estimates of population: the dress rehearsal census of 1988. U.S. Bureau of the Census, *1990 Annual Research Conference Proceedings*. 326-361.
- MULRY, M.H., and SPENCER, B.D. (1993). Accuracy of the 1990 census and undercount adjustments. *Journal of the American Statistical Association*. 88, 1080-1091.
- NATHAN, G. (1967). Outcome probabilities for a record matching process with complete invariant information. *Journal of the American Statistical Association*. 62, 454-469.
- S.I.S. (1994). 1990 Census of Population Response Reliability Survey. *State Institute of Statistics Publication No. 1688*, Ankara, 65.
- SRINIVASAN, S.K., and MUTHIAH, S. A. (1968). Problems of matching births identified from two independent sources. *The Journal of Family Welfare*. 14, 13-22.
- TEPPING, B.J. (1968). A model for optimal linkage of records. *Journal of the American Statistical Association*. 63, 1321-1332.
- WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*. 81, 338-346.

A Hierarchical Model for the Analysis of Local Census Undercount in Italy

D. COCCHI, E. FABRIZI and C. TRIVISANO¹

ABSTRACT

Census counts are known to be inexact based on comparisons of Census and Post Enumeration Survey (PES) figures. In Italy, the role of municipal administrations is crucial for both Census and PES field operations. In this paper we analyze the impact of municipality on Italian Census undercount rates by modeling data from the PES as well as from other sources using Poisson regression trees and hierarchical Poisson models. The Poisson regression trees cluster municipalities into homogeneous groups. The hierarchical Poisson models can be considered as tools for Small Area estimation.

KEY WORDS: Census undercount; Post enumeration survey; Bayesian hierarchical modelling; Gamma-Poisson regression models; Poisson regression trees.

1. INTRODUCTION

The Italian Population Census takes place every ten years and represents the most important institutional duty of the Italian National Institute of Statistics (ISTAT) (The work leading to this paper has been developed just before 2001 Italian Census and the subsequent PES. The results have been considered in performing the 2001 PES). In order to carry out the Census, ISTAT relies on municipal administrations who are responsible for all field operations (training of interviewers, planning of interviews, data gathering and basic data processing). During Census operations, each municipality works independently from the others under ISTAT supervision. The accuracy of the Census results therefore differ considerably from one municipality to another, even if contiguous. In Italy, the geographical area of a municipal borough is sub-divided into Census Enumeration Areas (EAs), which are assigned to a single interviewer during Census operations. The EAs differ in terms of shape, structure and difficulty of enumeration, as well as interviewer. It is likely that the undercount rate varies substantially among EAs within the same municipality.

After the 1991 Population Census, ISTAT conducted a Post Enumeration Survey (PES) to measure the phenomenon of undercount. Population Census counts are known to be generally incorrect because of missed, multiple and misplaced enumeration. Missed enumeration is the most important inaccuracy and typically yields a net population undercount that may vary geographically and between different social groups, and impacts the determination of the relative sizes of sub-populations (Abbate, Masselli, Signore 1993). Field operations of the PES were carried out by the sampled municipalities themselves. The 1991 Italian PES data have been analyzed by Abbate, Masselli and Signore (1993), who estimate the overall national undercount rate by means of a Lincoln-Petersen model (see Wolter 1986) using post-strata of municipalities based on large

geographical areas (North, Center, South). Working on the same data, Fortini (1994) estimates the overall national undercount by means of latent class models.

Instead of estimating the undercount rate for the whole country or smaller domains, we propose models designed to explain the variation in undercount rate at the municipal level. The availability of factors accounting for the size of the net undercount may be a basis for creating homogeneous groups of municipalities, for planning a more efficient stratification in future Post Enumeration Surveys. Moreover, knowledge of those flaws in municipal organization which significantly influence the undercount may provide guidelines for actions designed to reduce its size.

Contributions which use disaggregated PES data are present in the literature. Alho, Mulry, Wurdeman and Kim (1993) consider a logistic regression model for the individual (household) probability of being censused. In keeping with Moura and Holt (1999), their model could be extended to include municipality or other group effects. We are in fact aware that our choice of modelling municipal data is not the same as the analysis of household level records, since many features determining individual propensity to be caught by the Census average out when dealing with aggregated data. A comprehensive analysis based on individual records is not feasible in the Italian case, since there were very few questions for individuals included in the 1991 PES schedule. Similarly, the 1991 PES provides very little auxiliary information on the EAs, with the consequence that models based on EA undercounts cannot be proposed.

Our analysis is based on combining different data sources. The auxiliary information comes from the above-mentioned 1991 PES, two studies on the statistical quality of municipalities conducted by ISTAT during the early 90s (Di Pietro 1998, 1999) and demographic and social indicators obtained from the 1991 official Census results.

¹ D. Cocchi, E. Fabrizi and C. Trivisano, Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna, Italy.

We face the problem of how to make efficient use of the information obtained from the various data sources. We have in fact a large number of variables, most of which are categorical or polychotomous. Instead of using a variable selection algorithm, we have chosen to build homogeneous groups of municipalities which are then introduced into the model by means of a design matrix for the random effects. These groups are constructed using Poisson regression trees (Therneau and Atkinson 1997). This hierarchical usage of information provides a natural basis for the design of strata of geographically non-contiguous municipalities.

Few EAs are re-censused within each sampled municipality in the PES; the average EAs sampling rate is 0.001. This is a typical Small Area setting where direct estimates of the municipal undercount rate are unreliable and ought to be replaced by synthetic or composite estimates based on a suitable model. The phenomenon of undercount is rare. Our data consist of counts and may show a large overdispersion with respect to a Poisson distributional assumption. We suggest the use of hierarchical Poisson regression models to manage overdispersion.

The hierarchical models here adopted manage explicitly overdispersion due to municipal heterogeneity. A further extra Poisson variability source is due to heterogeneity within municipalities, because of clustering of missed enumeration within EAs, or of clustering due to missed enumerations of individuals in the same family. This kind of overdispersion is not explicitly treated in the models.

We adopt a full Bayesian approach for specification and estimation purposes and base the solution of the models on Markov chain Monte Carlo simulation methods. Within this hierarchical framework, we deal with overdispersion by imposing a Gamma distribution on the rate of the first level Poisson distribution, thus marginally obtaining a Negative Binomial. Moreover, conditionally on the hyperparameters, the proposed model features posterior linearity and the corresponding posterior means for the municipal undercount rates are linear composite estimators. Thus, the amount of smoothing depends on how much information is provided by each municipal sample in the PES.

Our results show that the municipality stratification employed in designing the 1991 PES (based on geographical area and population size) can be improved, since the undercount rate is shown to be largely independent of geographical area. On the contrary, variables describing the statistical efficiency of local administrations are useful in discriminating between the different degrees of undercount among municipalities of similar size and demographic structure. Whilst leaving the design of the PES unchanged, our results may provide useful guidance when performing data analysis.

The present paper is organized as follows. Section 2 describes the basic features of the PES and of the other data sources we have taken into consideration. Section 3 looks at the Poisson regression trees used to build homogeneous groups of municipalities. In section 4 we introduce the

hierarchical Poisson regression models, while empirical results and model comparisons are discussed in section 5.

2. THE PES DATA AND AUXILIARY INFORMATION

2.1 The Italian Post Enumeration Survey

The 1991 Italian Population Census took place on October 20th. The subsequent Post Enumeration Survey, based on a two stage stratified sampling design, was carried out a few weeks later. Municipalities constitute the primary units, whereas the secondary ones are represented by the Census EAs. An EA is the smallest area into which the municipal territory is partitioned for Census operations; each EA is assigned to a single interviewer.

The primary sampling units were stratified according to geographical area (North-West, North-East, Center, South, Islands) and demographic size (7 classes for the municipalities below 350,000 inhabitants), producing 35 strata. Within each stratum the sampled municipalities were selected without replacement and with probability proportional to their demographic size. The 10 municipalities with more than 350,000 inhabitants have been included in the sample as self-representative units. The secondary sampling units were selected with equal probabilities by systematic sampling. The final PES sample contains 85 municipalities and 638 EAs (out of a national total of 8,095 municipalities and 64,000 EAs) with a national design based estimate of 1.24% (Abbate, Masselli and Signore 1993).

The PES forms were filled out during face to face interviews and contained just a few simple questions. The characteristics of the sampled households are limited to the number and gender of household members. Other PES questions were designed to facilitate record linkage with the Census result, and therefore to reduce both misplaced enumeration and other non sampling errors in the evaluation of undercount (see Fortini 1994 for details).

2.2 The Surveys of the Statistical Quality of Municipalities

A data set on the statistical quality of Italian municipalities was constructed by ISTAT (see Di Pietro 1998, 1999). It integrates different sources: information from 1991 Census performance records, municipal population registers and Interior Ministry data. This data set contains also the results of three administrative surveys, conducted during the 90s, carried out in order to evaluate the performance of municipalities with regard to their commitments to ISTAT. The first survey is about the computerization of municipal Statistics Bureaus. The second survey, known with the acronym POSAS, is a post-Census survey of the demographic registers of the resident population, classified by year of birth, age and civil status. The third survey, known with the acronym ISCAN, regards the

appropriateness of registrations on the municipal population registers list. These surveys provide data for all Italian municipalities.

From this data set we selected a subset of variables related to the municipal activity at the time of the 1991 Census:

- a) the percentage of noncoded fields of the Census household forms which had to be filled out, after the interview of the households, by the municipal Statistics Bureaus (PERCOD);
- b) the ratio of the population temporarily abroad to the population present at the 1991 Census (PERCEST);
- c) the ratio of the difference between the 1991 Census and population registers counts to the 1991 Census counts (PERDIFF);
- d) the time needed to update municipal demographic registers on the basis of 1991 Census results (IND01);
- e) delay in street name updating (IND11).

2.3 Demographic Variables

We also consider a set of demographic ratios from the 1991 Census results. In particular, we use the percentages of "single member" and "more than one family" households, and sex ratios (males/females) in the municipality. The municipal resident population – resulting from the uncorrected 1991 Census counts – is also a very important variable. The number of EAs sampled in each municipality for the PES is a further signal of the municipality importance.

3. POISSON REGRESSION TREES

The available data sources provide us with a large number of auxiliary variables, many of which are categorical or polychotomous. Before we fit the hierarchical models, we group municipalities with homogeneous household undercount rates using Poisson binary regression trees. Groups based on trees are included as factors in the models described in the next section. Our principal aim is to check the effectiveness of traditional stratifications, improving them *ex post* by hierarchical models with suitable covariates and to verify how they differ from comparable results based on optimal groupings.

The conditional regression models are based on the canonical logarithmic link. The splitting criterion is based on the usual deviance statistic (Therneau and Atkinson 1997):

$$\text{Deviance}_{\text{parent}} - (\text{Deviance}_{\text{child, left}} + \text{Deviance}_{\text{child, right}})$$

The basic idea for building a tree is to begin with a large tree T_0 constructed using a naive and mild stopping rule (as the minimum number of observations in the final nodes of the tree) and then to select the right-sized tree among the

sub-trees of T_0 by pruning. The established methodology for pruning trees is cost-complexity pruning, first introduced by Breiman, Friedman, Olshen and Stone (1984). Let D_T be the deviance of a subtree T of T_0 , $\text{size}(T)$ the number of terminal nodes of T and $\alpha > 0$ a cost-complexity parameter for defining the cost-complexity measure:

$$D_T(\alpha) = D_T + \alpha \text{size}(T) \quad (1)$$

For a specified α the tree $T(\alpha)$ that minimizes (1) can be found. It can be shown (Breiman *et al.* 1984) that a nested family of subtrees $\{T_0, T_1, \dots, T_k, \dots, T_{\text{root}}\}$ of T_0 exists such that each tree is optimal for a range of values of α .

The problem is now reduced to selecting one of these subtrees. The selection is carried out in order to minimize the prediction error defined as the deviance contribution for a new observation. To estimate the prediction error, the availability of an independent sample would be in principle the best option, but since it is advisable to use all data to "instruct" the tree in the best possible way, a cross-validation method is used. Usually, the tree T_{k_0} with the minimum estimated prediction error is selected. Here we use a more severe pruning rule which consists in selecting the smallest tree with an estimated prediction error not larger than the estimated prediction error of T_{k_0} plus its standard error. This pruning rule, known as the "1 SE rule" (Breiman *et al.* 1984), is adopted in order to avoid model overfitting.

Since the cross-validation of Poisson regression trees may give, in some nodes, infinite values for the deviance statistic, we use Bayesian shrinkage estimators of the true rates, based on a simple Poisson-Gamma model, as suggested in Thernau and Atkinson (1997).

We built three different trees based on different starting subsets of auxiliary variables.

Tree 1 (shown in Figure 1) is based on demographic variables only. The first split separates municipalities with population less than 100,100 from those with more than 100,100. This splitting value is almost coincident with the 100,000 demarcation value used in the stratification of municipalities for the 1991 PES. The second split isolates a sub-sample of small municipalities for which less than 4 EAs were sampled in the PES. A further split is made on the basis of the sex ratio.

Tree 2 (Figure 2) is based exclusively on variables concerning the quality of the statistical performance of municipalities. The first split is based on the timing in correcting demographic registers (IND01): those municipalities that were quickest in performing this activity have the lowest undercount rates. Lower level splits highlight the problem of people temporarily abroad (PERCEST) which in areas characterized by massive emigration may lead to serious undercounting of the municipal population and errors in the book-keeping of demographic registers (PERDIFF). In this tree, one half of the sample is classified in a single node which is likely to contain residual heterogeneity.

Tree 3 (Figure 3) is based on both demographic and quality variables. The first split is based on the municipal population exactly as was the case in Tree 1. Subsequently, the subset of municipalities with less than 100,100

inhabitants is split into small and middle sized municipalities at a threshold of 13,200. The quality variable included in this tree consists of timing in correcting demographic registers (IND01).

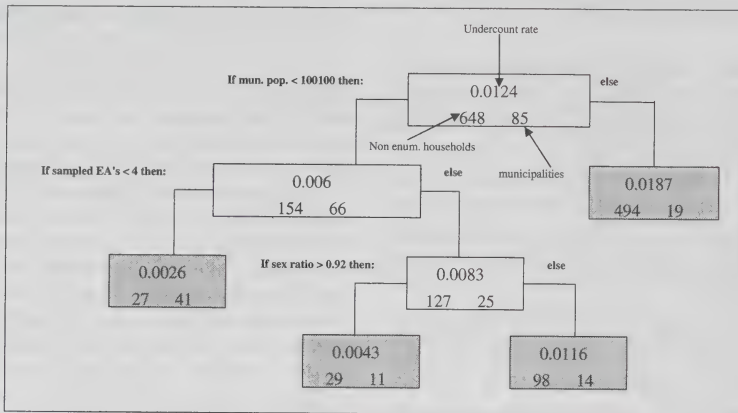


Figure 1. Tree 1 based on demographic variables.

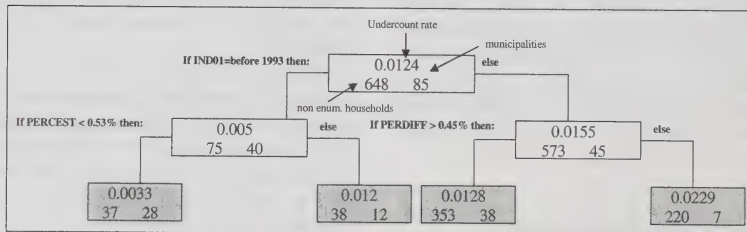


Figure 2. Tree 2 based on municipal statistical quality variables.

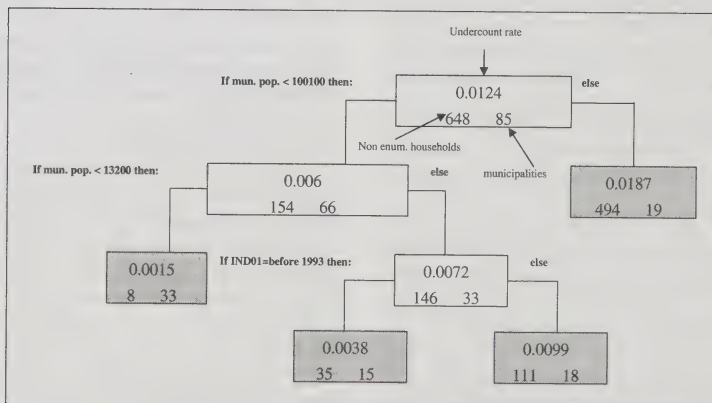


Figure 3. Tree 3 based on demographic and quality variables.

4. HIERARCHICAL POISSON-GAMMA MODELS

We denote the observed number of not enumerated households in each municipal sample with $y_i (i = 1, \dots, 85)$. As an initial approximation, these counts can be modeled using a Poisson distribution:

$$y_i | \delta_i, e_i \sim \text{Pois}(\delta_i e_i) \quad (2)$$

where δ_i represents the rate of undercount to be estimated and e_i is given by the number of households in the sampled EAs within the municipality. Dependency on a set of explanatory variables is expressed by means of a canonical log-linear link:

$$\ln(\delta_i e_i) = X_i \beta + Z_i \xi \quad (3)$$

where Z_i is the i -th row of a categorical design matrix introduced for modelling group effects. Each X_i is a p -vector of explanatory variables associated with the i -th municipality and β and ξ are the regression parameters.

The occurrence of failure to enumerate is relatively rare when compared to the number of observed households. For this reason, the data may show strong overdispersion. Overdispersion can be managed by hierarchically modelling the parameters δ_i in (2). If the δ_i are $\text{Gamma}(\alpha, \nu)$ distributed, the Negative Binomial distribution is marginally obtained for y_i by integrating out the parameters δ_i : i.e. $y_i | \alpha, \nu, e_i \sim \text{NegBin}(\alpha, \nu/(\nu + e_i))$ with moments:

$$E(y_i | \alpha, e_i, \nu) = \frac{\alpha e_i}{\nu}, \quad V(y_i | \alpha, e_i, \nu) = \frac{\alpha e_i (\nu + e_i)}{\nu^2}$$

(see Lawless 1987).

Instead of the parameterization above, we adopt the parameterization of the Gamma distribution at the second level of the hierarchy according to the proposal made by Christiansen and Morris (1997). When assuming

$$\delta_i | \lambda_i, \zeta \sim \text{Gamma}(\zeta, \zeta/\lambda_i) \quad (4)$$

with moments $E(\delta_i | \lambda_i, \zeta) = \lambda_i$ and $V(\delta_i | \lambda_i, \zeta) = \lambda_i^2/\zeta$, we have

$$y_i | e_i, \lambda_i, \zeta \sim \text{NegBin}\left(\zeta, \frac{\zeta/\lambda_i}{\zeta/\lambda_i + e_i}\right),$$

where $V(y_i | e_i, \lambda_i, \zeta) - E(y_i | e_i, \lambda_i, \zeta) = e_i^2 \lambda_i^2/\zeta$. As ζ moves towards infinity, the variance of the Negative Binomial converges towards that of the Poisson (the variance of the Gamma in (4) tends towards 0), while small values of ζ point to high overdispersion.

From (4) it is immediate to see that:

$$E(\delta_i e_i | e_i, \lambda_i, \zeta) = \lambda_i e_i;$$

therefore the dependence assumption (3) is re-stated in terms of $\lambda_i e_i$:

$$\ln(\lambda_i e_i) = X_i \beta + Z_i \xi.$$

The prior (4) is conjugate to the likelihood defined by (2). Consequently one obtains

$$\delta_i | y_i, e_i, \lambda_i, \zeta \sim \text{Gamma}(y_i + \zeta, e_i + \zeta/\lambda_i)$$

from which it follows that

$$E(\delta_i | y_i, e_i, \lambda_i, \zeta) = (1 - B_i) r_i + B_i \lambda_i \quad (5)$$

where $r_i = y_i/e_i$ and $B_i = \zeta/(\zeta + e_i \lambda_i)$.

Each posterior mean (5) can be seen as a composite Small Area estimator where both the direct and the synthetic components are weighted according to the information available from the sample.

From (5) we note that the posterior mean of the distribution of the rate parameters δ_i is a linear combination of the observed undercount rate r_i and the prior mean λ_i . In other words, the model features posterior linearity. The two terms in (5) are weighted according to B_i , which varies between 0 and 1. The larger the B_i , the more the prior means λ_i (synthetic estimators) receive weight and the model estimates gain in importance compared with the observed rates. We note that each B_i is inversely proportional to the $e_i \lambda_i$, expressing the amount of information provided by the sample of each domain.

To complete the full Bayesian specification of the model we assign a distribution to the third level parameters ζ, β, ξ . According to an approximate non-informative criterion, we introduce proper, but flat, prior distributions. In particular we assume that:

$$\beta_j \stackrel{\text{iid}}{\sim} N(0, 100), \quad j = 1, \dots, p \quad (6)$$

$$\xi_k \stackrel{\text{ind}}{\sim} N(k \bar{u}_k, \frac{1}{\tau \bar{n}_k}), \quad k = 1, \dots, q \quad (7)$$

where \bar{u}_k is the average undercount in the k -th group and \bar{n}_k is the average number of sampled households in the municipalities of the same group. Priors (7), associated to group effects, are therefore centered on groups means and their precision is proportional to the group size. They are built to be weakly informative for improving the stability and convergence properties of the model. Priors for regression coefficients (6) associated to the remaining regressors are centered in 0. For the overdispersion parameter ζ we select the prior

$$\zeta \sim 1,000^+ \text{Gamma}(0.001, 1) \quad (8)$$

following the suggestion given by Christiansen and Morris (1997). Note that the first two prior moments of (8) are $E(\zeta) = 1$ and $V(\zeta) = 1,000$; thus the prior is very diffuse and characterized by high positive skewness.

At the fourth level of the hierarchy we specify the following priors:

$$k \sim N(0, 100) \quad (9)$$

$$\tau \sim \text{Gamma}(0.001, 0.001). \quad (10)$$

which are both designed to have a very mild impact on posterior inferences.

We compute the posterior distributions of $(\delta_i | y_i, e_i)$ by using Markov chain Monte Carlo (McMC) sampling algorithms. For these calculations we use the software BUGS (Spiegelhalter, Thomas, Best and Gilks 1995), which is based on Gibbs sampling. Since the solution of models involving discrete distributions is computationally very demanding, we specify the prior distributions (6) – (10) by selecting simple well known functional forms, as Normal and Gamma, that facilitate fast computations. We examined the sensitivity of the posterior means in (6) – (10), and we did not find any substantial changes in the posterior means. Hence, these priors can be considered noninformative. For the convergence assessment we consider the multiple chain approach suggested by Gelman and Rubin (1992), running three different chains with well separated starting points for each model. The visual inspection of the chains path and the modified Gelman and Rubin statistic (Brooks and Gelman 1998) are considered as basic convergence assessment tools. We run 10,000 iterations for each chain, discarding on average a conservative “burn in” of 3,000, thus yielding an approximate 20,000 draws from the posterior of each model.

5. MODEL COMPARISON AND DISCUSSION OF EMPIRICAL RESULTS

We estimated a variety of models for different definitions of the matrixes of regressors X and Z . As regards the design matrix Z we consider seven different cases, in which municipalities are grouped using either traditional stratification criteria (geographical area and demographic size) or the results of the partitioning techniques discussed in section 3. They are: a) geographical area (North, Center, South and Islands), b) demographic size classes only, c) demographic classes by geographical area, d) demographic size classes and geographical areas, e) Tree 1 (based on demographic variables), f) Tree 2 (based on quality variables), g) Tree 3 (based on both quality and demographic variables). Two kinds of variables may be proposed in matrix X : the quality variables of section 2.2 and the demographic variables of section 2.3. Matrix X has therefore three different possible compositions: I) quality variables only, II) demographic variables only, III) both quality and demographic variables. By matching the different definitions of X and Z , twenty-eight different models have been estimated. In this way we do not perform variable selection procedures, rather we introduce alternative blocks of variables.

The quantity commonly used for comparing models within the Bayesian framework is the Bayes factor (BF). A large sample approximation of $-2\ln(BF)$ is given by

$$\Delta BIC = -2\ln \left[\frac{\sup_{M_0} f(y | \theta_0)}{\sup_{M_k} f(y | \theta_k)} \right] - (p_k - p_0) \ln n \quad (11)$$

(see Schwarz 1978) which, moreover, makes no reference to the prior assumptions. We note that in (11) the M_k ($k = 1, \dots, K$) index the set of competing models and θ_k is the p_k dimensional parameter indexing the likelihood associated to each model. The null model against which all the others are compared is the one with the only intercept, and is denoted by M_0 . Positive and large values of (11) support model M_k .

The complexity penalization in (11) depends on the size of the subset of third level parameters; that is, all models are compared as if they were non hierarchical. Since they share a similar hierarchical structure, this operational modification of the standard BIC criterion does not alter the results of model comparison summarized in Table 1.

We note that those models where group effects are based on geographical area perform very poorly (row 1), and the same happens when the geographical area is combined with the demographic size of the municipalities (rows 3 and 4). This is rather surprising, since geographical areas are employed in designing the stratification of the PES sample, and the efficiency of administrations, together with other social and economic indicators, are currently supposed to be clustered with respect to Italy's large geographical subdivisions (North, Center, South). This outcome may be ascribed to the predominant role that the specific organization of each municipality plays in determining the efficiency of Census operations within its territory.

Models with tree-based group effects (rows 5-7) clearly perform better than models with group effects based on ISTAT traditional stratification criteria (rows 1-4). The only exception to this behavior are those models relying on Tree 2 (row 5), which perform rather poorly when demographic size and other demographic variables are not included. In fact, the municipal population can be thought of as a proxy of municipal organizational complexity. It seems that quality variables are powerful in discriminating the level of undercount among municipalities with similar demographic features, but have little relevance when the effect of a different degree of organizational complexity is not accounted for by introducing a variable of demographic size. We point out that adding a design matrix Z based on Poisson regression trees grouping of municipalities allows us to model non linear relations between the undercount and the predictors.

Actually, the models based on Tree 3 provide the best performance. A number of comments about the model with maximum ΔBIC follow. This model uses demographic and quality variables as regressors. The adequacy of the selected model is assessed by means of posterior predictive checks. In particular the general purpose goodness-of-fit discrepancy measure proposed by Brooks, Catchpole and Morgan (2000) as a suitable tool for rare occurrences as census undercounts:

$$D(y;\theta) = \sum_i \left(\sqrt{y_i} - \sqrt{E\text{xp}_i} \right)^2, \tag{12}$$

where $\text{Exp}_i = e_i E(\delta_i | y_i, e_i)$, is adopted. The associated 0.46 tail area probability highlights a good fit for the selected model.

The set of models has been estimated again after eliminating the greatest municipality, which is potentially an influential case. Again, the model based on Tree 3 with demographic and quality variables as regressors has been selected using the criterion (11). This model shows a good fit (the Bayesian p -value associated to the discrepancy measure (12) is equal to 0.51). Moreover, composite estimates do not change much when compared with those obtained with the whole sample.

In order to check model fitting, in Figure 4, composite estimates against direct estimates of the number of not enumerated household in each municipality are plotted (the values of the largest 10 municipalities are reported with a different scale). The composite estimates are $w_i e_i E(\delta_i | y_i, e_i)$, while the direct estimates are $w_i y_i$, w_i being the expansion factor due to EA sampling in each municipality. Composite estimates are posterior expectations of first level parameters and, conditionally on the hyperparameters, are composite estimates in which the model predictions represented by the λ_i receive little weight when there is sound sampling evidence. From (5) we know that this weighting process is ruled by the municipal shrinkage factors B_i . They weight the direct estimates y_i/e_i in proportion to $e_i \lambda_i$, i.e. the number of not enumerated households within the municipal sample predicted by the model.

Table 1
 ΔBIC of the estimated models compared with the reference model M_0

		Variables in the models			
Group Effects	Area	Only group effects	Group eff. + quality vars	Group eff. + demographic vars	Group eff. + quality and demographic vars
	Classes of Mun. Pop.	-4.22	-0.39	18.52	23.32
	Area* Mun. Pop. Classes	15.34	17.87	17.32	20.09
	Area + Mun. Pop. Classes	2.08	6.13	4.91	8.45
	Tree2 (quality vars)	9.68	13.20	13.74	17.83
	Tree1 (demographic vars)	11.81	8.34	23.48	26.15
	Tree3 (quality + demographic vars)	35.14	35.37	32.28	35.53
		38.89	35.76	41.12	41.45

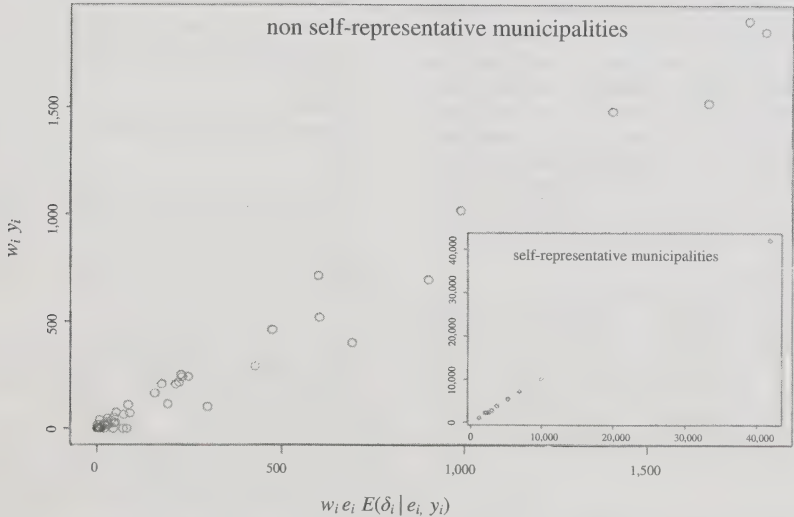


Figure 4. Composite estimates against direct estimates of the number of not enumerated households in each municipality.

For municipalities with resident population of up to 10,000 (this value is relatively close to the splitting value 13,200 of Tree 3) in almost all cases we have B_i values that are very close to 1; this means that, for small municipalities, the role of the model component in the determination of the composite estimate is overwhelming. In Figure 5 composite estimates (×) and their 95% credibility intervals are plotted against direct estimates.

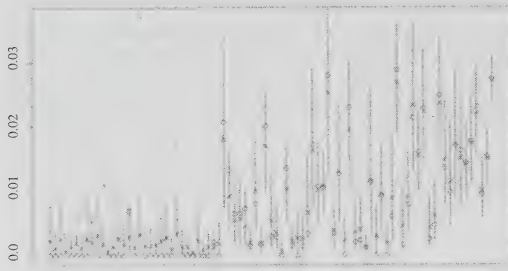


Figure 5. Composite estimates (×) and their 95% credibility intervals; (○) direct estimates. Municipalities are sorted by demographic size.

The width of the credibility intervals depends on the undercount level and, as should be expected, is large when the size of the sample within the municipality is small. Composite estimates associated with large credibility intervals are also characterized by large shrinkage factors, as a consequence of the scarce sample information. Large intervals for some middle-sized municipalities can be justified with the fact that they are under-sampled with respect to their size.

In small municipalities, where Census is conducted more easily, the undercount is generally very small. The undercount estimate is difficult since very few EAs are currently sampled from each of the small municipalities, often providing no evidence of undercount. In such cases, the composite estimate essentially consists in the model based component. Therefore, for the next PES, given the overall sample size, our suggestion is not to insist in sampling a great number of small municipalities, but to redirect sampling towards middle-sized municipalities, which are more heterogeneous. Moreover, the number of EAs to sample in the selected small municipalities ought to be increased.

The results of this work, which considers for the first time a criterion for grouping together municipalities according to their performance in statistical operations, confirm that an improvement may be reached for future similar surveys by modifying the stratified sampling design and by modelling undercount by means of the covariates mimicking the difficulties of the municipality behaviour in conducting censuses.

ACKNOWLEDGEMENTS

We would like to thank Angela Ferruzza, Marco Fortini, Aldo Orasi and Fernanda Panizon of the ISTAT team working on the 2001 Census and PES, together with Mariella Dimitri and Ersilia Di Pietro, of the ISTAT group working on surveys of statistical performance of municipalities, for their useful suggestions and continuous assistance.

The work has been partially funded by the (1999-2000) "Quality of total and partial surveys" Research Project grant from the University of Bologna (60%).

The PES data set and the archives containing the data on municipalities have been made available thanks to a special agreement between ISTAT and the Department of Statistics of the University of Bologna.

We would like to thank Francesca Bruno and Loredana Di Consiglio for their invaluable contribution in preparing the basic data sets, and Meri Raggi for her constant support and her discussion of the subjects of this research.

We thank the Editor, an Associate Editor and two anonymous referees for comments and suggestions which helped us in revising and improving the manuscript.

REFERENCES

- ABBATE, C., MASSELLI, M. and SIGNORE M. (1993). A combined post-enumeration survey for the 1991 Italian population and industrial censuses. *Bulletin of the International Statistical Institute, Firenze, 48th Session*. Tome LV, 2, 159-173.
- ALHO, J.M., MULRY, M.H., WURDEMAN, K. and KIM, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*. 88, 1130-1136.
- BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J., (1984). *Classification and Regression Trees*. Wadsworth, California.
- BROOKS, S.P., CATCHPOLE, E.A. and MORGAN, B.J.T. (2000). Bayesian animal survival estimation. *Statistical Science*. 15, 357-276.
- BROOKS, S.P., and GELMAN, A. (1998). Alternative methods for monitoring convergence of iterative simulation. *Journal of Computational and Graphical Statistics*. 7, 434-455.
- CHRISTIANSEN, C.L., and MORRIS, C. (1997). Hierarchical Poisson regression models. *Journal of the American Statistical Association*. 92, 618-632.
- DI PIETRO, E. (1998). Anagrafi comunali: funzione statistica e livello di informatizzazione. *Atti Della Quarta Conferenza Nazionale di Statistica*. Tomo 1 - Sessioni Plenarie, Workshop: Il progetto anagrafi. Roma 11-13 novembre.
- DI PIETRO, E. (1999). Anagrafe informatizzata e Censimenti demografici: dal censimento tradizionale al censimento basato sugli Archivi. *Società Italiana di Statistica: Atti Del Convegno "Verso i Censimenti del 2000"*. Udine 7-9 giugno. 169-182.

- FORTINI, M. (1994). Un'applicazione del modello a classi latenti per l'analisi dell'errore di copertura del XIII censimento della popolazione. *Atti della XXXVII Riunione Scientifica della Società Italiana di Statistica*. San Remo 6-8 Aprile. 2, 423-430.
- GELMAN, A., and RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequence. *Statistical Science*. 7, 457-72.
- LAWLESS, J.F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*. 15, 209-225.
- MOURA, F.A.S., and HOLT, D. (1999). Small area estimation using multilevel models. *Survey Methodology*. 25, 73-80.
- SCHWARTZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics*. 6, 461-464.
- SPIEGELHALTER, D.J., THOMAS, A., BEST, N. and GILKS, W.R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50*. Technical Report, Medical Research Council biostatistics Unit, Institute of Public Health, Cambridge University.
- THERNEAU, T.M., and ATKINSON, E.J. (1997). *An Introduction to Recursive Partitioning Using the RPART Routines*. Technical report, Mayo Foundation.
- WOLTER, K. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*. 81, 338-346.

Estimation of a Measure of Disclosure Risk for Survey Microdata Under Unequal Probability Sampling

C.J. SKINNER and R.G. CARTER¹

ABSTRACT

Skinner and Elliot (2002) proposed a simple measure of disclosure risk for survey microdata and showed how to estimate this measure under sampling with equal probabilities. In this paper we show how their results on point estimation and variance estimation may be extended to handle unequal probability sampling. Our approach assumes a Poisson sampling design. Comments are made about the possible impact of departures from this assumption.

KEY WORDS: Confidentiality protection; Finite population inference; Poisson sampling; Statistical disclosure control; Uniqueness.

1. INTRODUCTION

Microdata files of survey data may be of great analytic value to researchers. When deciding whether and how to make such files available, agencies conducting surveys need to protect against risks of possible statistical disclosure (Willenborg and de Waal 2001). Skinner and Elliot (2002, abbreviated henceforth to SE) proposed a simple measure of statistical disclosure risk for survey microdata, for use as evidence to inform such decisions. They showed that this measure may be estimated simply under sampling with equal probabilities. In this paper we show how their results may be extended to handle unequal probability sampling.

The measure is introduced in section 2. Point estimation and variance estimation for the measure are considered in sections 3 and 4 respectively. See SE for the relation of this measure to the literature on statistical disclosure risk.

2. THE MEASURE OF DISCLOSURE RISK

We consider the possible release of a microdata file consisting of a set of records for units (*e.g.*, individuals or households) in a sample s , selected by probability sampling from a population U . Each record consists of a vector of values of a specified set of variables for the given unit. Following a standard approach to disclosure risk assessment (*e.g.*, Bethlehem, Keller and Pannekoek 1990), we suppose that an intruder attempts to match the microdata records to known population units using a specified subset of variables. We assume that these 'identifying variables' are categorical and that the possible combinations of their values define the categories 1, ..., J of a variable X . (J will usually be very large).

We suppose that the intruder is able to determine the value of X for a population unit with known identity and

that the intruder 'claims' that a microdata record has been identified if and only if this value matches the value of X recorded in the microdata for *just one* microdata record. Assuming (a) that the population unit with known identity is randomly drawn from U with equal probabilities and (b) that the value of X for this unit is measured in the same way that X is measured in the microdata, the probability that the intruder's claim is correct is:

$$\begin{aligned}\theta &= \Pr(\text{correct match} \mid \text{unique match}) \\ &= \sum_{j=1}^J I(f_j = 1) / \sum_{j=1}^J F_j I(f_j = 1),\end{aligned}$$

where f_j and F_j are the frequencies of units in s and U respectively, for which $X = j$ and where $I(\cdot)$ is the indicator function ($I(A) = 1$ if A is true, $I(A) = 0$ otherwise). The numerator of θ is the number of microdata records which are unique in the microdata with respect to X and the denominator of θ is the number of units in the population which share the same value of X with any of these records.

The quantity, θ , is the measure of disclosure risk considered in this paper. To protect against disclosure, θ might be estimated under alternative forms of microdata release (implying alternative specifications of X) and a form of release chosen so that θ is inferred to be acceptably small. A sensitivity analysis will usually be required in which the specification of X is varied not only according to the form of release but also to allow for alternative plausible forms of external information which an intruder might hold about known population units. For example, one might consider both an intruder with access only to publicly available information, such as the visible characteristics of an individual, and an intruder with access to a private database held by an organisation.

¹ C.J. Skinner, University of Southampton, Southampton, United Kingdom, S017 1BJ and R.G. Carter, Statistics Canada, B-2 Jean Talon Building, Ottawa, Ontario, K1A 0T6.

3. ESTIMATION OF θ

We suppose that the data consist of the values of X for the sample units. Hence, the sample frequencies f_j are known but the population frequencies F_j are unknown ($j = 1, \dots, J$). The ‘parameter’ of interest, θ , is also unknown and must be estimated. We adopt a design-based approach to inference in which the f_j are random and the F_j are fixed. As discussed by SE, the ‘parameter’, θ , therefore depends on s , unlike standard finite population parameters considered in survey sampling.

SE motivate a point estimator of θ by a resampling argument, which may be generalised to the case of unequal probability sampling, as follows.

Repeat the following steps K times.

Step 1: remove a single unit i from the microdata sample s with probability

$$\alpha_i = \pi_i^{-1} / \sum_s \pi_i^{-1},$$

where π_i is the (first-order) inclusion probability of unit i ;

Step 2: copy the removed unit back into the sample with probability π_i ;

Step 3: record whether the removed unit matches a unique record in the microdata and whether this match is correct.

The idea is that Step 1 mimics the intruder’s (equal probability) selection of a unit from U (using the inverse sampling idea of Hinkins, Oh and Scheuren 1997). Step 2 mimics the inclusion of that unit in s . The estimator of θ is the empirical proportion of unique matches which are correct. Following the argument of SE, this estimator converges almost surely as $K \rightarrow \infty$ to

$$\begin{aligned} \hat{\theta} &= \frac{\sum_{s^{(1)}} \Pr(\text{unit } i \text{ removed and then copied back})}{\left[\sum_{s^{(1)}} \Pr(\text{unit } i \text{ removed and then copied back}) \right. \\ &\quad \left. + \sum_{s^{(2)}} \Pr(\text{unit } i \text{ removed and then not copied back}) \right]} \\ &= \sum_{s^{(1)}} \alpha_i \pi_i / \left[\sum_{s^{(1)}} \alpha_i \pi_i + \sum_{s^{(2)}} \alpha_i (1 - \pi_i) \right] \\ &= n^{(1)} / \left[n^{(1)} + \sum_{s^{(2)}} (\pi_i^{-1} - 1) \right], \end{aligned} \quad (1)$$

where $s^{(1)}$ is the subsample of unique units in s , $s^{(2)}$ is the subsample of units which occur in pairs and $n^{(1)} = \sum_j I(f_j = 1)$ is the size of $s^{(1)}$. In the case of equal probability sampling with $\pi_i = \pi$, $\hat{\theta}$ reduces to $n^{(1)} / [n^{(1)} + 2n^{(2)}(\pi^{-1} - 1)]$, where $2n^{(2)} = 2\sum_j I(f_j = 2)$ is the size of $s^{(2)}$, as in SE.

We are interested in $\hat{\theta}$, defined in (1), as an estimator of θ . SE show that $\hat{\theta}$ is consistent for θ in the equal probability sampling case. The basic steps of their argument may be generalised to the case of unequal probability sampling as follows. We may write

$$\theta = n^{(1)} / \left[n^{(1)} + \sum_j (F_j - 1) I(f_j = 1) \right]. \quad (2)$$

Hence, by comparing (1) and (2), $\hat{\theta}$ will be a ‘good’ estimator of θ if $\sum_{s^{(2)}} (\pi_i^{-1} - 1)$ is a ‘good’ estimator of $\sum_j (F_j - 1) I(f_j = 1)$. In fact, we prove in Appendix 1 that the latter estimator is unbiased, that is

$$E \left[\sum_{s^{(2)}} (\pi_i^{-1} - 1) \right] = E \left[\sum_j (F_j - 1) I(f_j = 1) \right], \quad (3)$$

under the assumption of Poisson sampling, that is where population units are sampled independently. Equation (3) generalizes Proposition 2 of SE. In the equal probability sampling case SE show how the result in equation (3) may be extended to prove consistency of $\hat{\theta}$ as an estimator of θ , using an asymptotic framework where $J \rightarrow \infty$ and under some regularity conditions, in particular that the F_j are bounded.

Having established the main unbiasedness result in (3), we conjecture that this consistency result will generalise to the case of unequal probability Poisson sampling, subject to additional weak conditions on the π_i , for example that the π_i are bounded above by a positive constant.

The Poisson sampling assumption generalises the Bernoulli sampling assumption in SE. They conclude that in practice $\hat{\theta}$ will remain approximately unbiased for θ under a number of other equal probability sampling designs including simple random sampling, (equal probability) systematic sampling or proportionate stratified simple random sampling. We suggest that in a similar way $\hat{\theta}$ will remain approximately unbiased for θ under corresponding unequal probability designs, *i.e.*, disproportionate stratified simple random sampling and unequal probability systematic sampling. We also suggest that it may be reasonable to allow for nonresponse in $\hat{\theta}$ if s is the set of respondents and if π_i^{-1} consists of a weight which may be interpreted as the reciprocal of the estimated probability of both being sampled and responding.

As discussed in SE, the form of sampling which seems to have the potential to lead to most bias in $\hat{\theta}$ as an estimator of θ in practice is multistage sampling, where the multistage units are strongly related with respect to X . For example, bias might be non-negligible when households form clusters within which all adult individuals are sampled, where the microdata includes individual-level records and where X is primarily determined by household-level variables. This might lead to a higher value of $n^{(2)}/n^{(1)}$ than expected under Poisson sampling and hence to underestimation of θ . Such an example is somewhat contrived, however, and we suspect that the bias of $\hat{\theta}$ as an estimator of θ will be modest in most typical social surveys.

4. VARIANCE ESTIMATION

SE present a linearization estimator of $\text{var}(\hat{\theta} - \theta)$, which depends on $n^{(1)}$ and $n^{(2)}$, like $\hat{\theta}$, as well as on $n^{(3)} = \sum_j I(f_j = 3)$, the number of values of X for which there are exactly three microdata records. We show in Appendix 2 that this variance estimator may be generalised, in the case of unequal probability Poisson sampling, to

$$\hat{v} = \hat{\theta}^2 \frac{\sum_{j=1}^J \left\{ I(f_j = 3)(\gamma_{1j}^2 - \gamma_{2j}) + I(f_j = 2)(\gamma_{1j}^2 + \gamma_{1j}) \right\}}{\left[n^{(1)} + \sum_j I(f_j = 2)\gamma_{1j} \right]^2} \quad (4)$$

where $\gamma_{1j} = \sum_s \beta_i$, $\gamma_{2j} = \sum_s \beta_i^2$, $\beta_i = \pi_i^{-1} - 1$ and $s_j = \{i \in s; X_i = j\}$, where X_i is the value of X for unit i .

Note that, in this notation, we may write

$$\hat{\theta} = n^{(1)} / \left[n^{(1)} + \sum_j I(f_j = 2)\gamma_{1j} \right].$$

As in the equal probability case, both $\hat{\theta}$ and \hat{v} can be computed straightforwardly from the values X_i and π_i for $i \in s$. The expression given above for \hat{v} reduces to the expression given in Proposition 3 of SE when $\pi_i = \pi$ for all $i \in s$.

The linearisation argument which gives \hat{v} assumes J is large. This seems a weak condition relative to the assumption of Poisson sampling. The linearisation variance estimator does not appear to generalise straightforwardly to other complex sampling designs. This is because the linearised form of $\hat{\theta} - \theta$ depends on the F_j and these cannot simply be replaced by consistent estimators. It also does not appear to be straightforward to apply replication methods to estimate the variance of $\hat{\theta} - \theta$, since θ is unknown and, as indicated by the simulation study in SE, the variance of θ may not be negligible in practice relative to the variance of $\hat{\theta}$.

5. CONCLUDING REMARKS

The estimated measure $\hat{\theta}$ considered in this paper may be used as evidence in assessing whether or not a proposed microdata file has an acceptable level of disclosure risk. The aim may be to ensure that $\hat{\theta}$ does not exceed some specified probability. To allow for sampling variation in $\hat{\theta}$ a more conservative procedure would be to require that the upper bound of a confidence interval for θ , say $\hat{\theta} + 2\hat{v}^{1/2}$, does not exceed the specified probability.

As well, $\hat{\theta}$ may be used to compare alternative strategies to control disclosure risk. For example, variables may be included in microdata files with more or less classification detail. Greater detail may enhance the value of the file for analysis but may also increase disclosure risk if the variable

could be used to match against external information. The estimated measure $\hat{\theta}$ could, therefore, be used to assess the relative risk resulting from different ways of collapsing the level of classification in specific identifying variables, including geography.

The measure may be estimated not only for the population as a whole, but also for subpopulations. Such a breakdown of the measure permits a more realistic assessment of the risk posed by intruders who target specific subpopulations. Such a targeted threat invalidates the basic assumption underlying the definition of whole population measure, θ , that the population unit with known identity is randomly drawn from U with equal probabilities. Separate estimation of the measure in different strata with different sampling fractions also provides a simple method of handling unequal probabilities of selection. This paper has shown how to allow for more general sources of unequal probability sampling in $\hat{\theta}$ and \hat{v} . More research is required to assess the robustness of these estimators to departures from Poisson sampling, especially multi-stage sampling.

A potential problem with estimating the measure separately by subpopulations is the impact of the reduction in sample size. SE found $\hat{\theta}$ to be stable in their numerical investigations, with a coefficient of variation never exceeding 6%. Their minimum sample size was, however, about 9,000 so further numerical work is needed to assess the stability of $\hat{\theta}$ for smaller sample sizes. The proposed variance estimation method provides some guidance for any specific case. Stability could, in principle, be improved by the use of model assumptions and one of us (CJS) is conducting further research on the limiting case of a small subpopulation, a single unit, extending θ to a record-level measure of risk analogous to that considered by Skinner and Holmes (1998).

APPENDIX 1

Proof of Equation (3)

Let $\beta_i = \pi_i^{-1} - 1$ and $U_j = \{i \in U; X_i = j\}$, $j = 1, \dots, J$, where X_i denotes the value of X for unit i . The size of U_j is F_j . Instead of labelling units in U by the single index i , consider the double index (jk) , $j = 1, \dots, J$, $k = 1, \dots, F_j$, so that, for example, $\pi_{(jk)}$ denotes the inclusion probability for the k -th unit in U_j and $\beta_{(jk)} = \pi_{(jk)}^{-1} - 1$. Under Poisson sampling the right side of (3) is

$$\begin{aligned} & E \left[\sum_j (F_j - 1) I(f_j = 1) \right] \\ &= \sum_{j=1}^J (F_j - 1) \sum_{k=1}^{F_j} \pi_{(jk)} \prod_{\substack{l=1 \\ l \neq k}}^{F_j} (1 - \pi_{(jl)}) \end{aligned} \quad (A.1)$$

and the left side of (3) is

$$\begin{aligned}
 & E\left[\sum_{i \in S^{(2)}} \beta_i\right] \\
 &= \sum_{j=1}^J \sum_{k=1}^{F_j} \sum_{\substack{\ell=1 \\ k < \ell}}^{F_j} \pi_{(jk)} \pi_{(j\ell)} \left[\prod_{\substack{m=1 \\ m \neq k, \ell}}^{F_j} (1 - \pi_{(jm)}) \right] [\beta_{(jk)} + \beta_{(j\ell)}] \\
 &= \sum_{j=1}^J \sum_{k=1}^{F_j} \sum_{\substack{\ell=1 \\ k \neq \ell}}^{F_j} \pi_{(jk)} \pi_{(j\ell)} \left[\prod_{\substack{m=1 \\ m \neq k, \ell}}^{F_j} (1 - \pi_{(jm)}) \right] \beta_{(jk)} \\
 &= \sum_{j=1}^J \sum_{k=1}^{F_j} \sum_{\substack{\ell=1 \\ k \neq \ell}}^{F_j} \pi_{(j\ell)} \left[\prod_{\substack{m=1 \\ m \neq \ell}}^{F_j} (1 - \pi_{(jm)}) \right] \\
 &= \sum_{j=1}^J (F_j - 1) \sum_{\ell=1}^{F_j} \pi_{(j\ell)} \left[\prod_{\substack{m=1 \\ m \neq \ell}}^{F_j} (1 - \pi_{(jm)}) \right]
 \end{aligned}$$

which is identical to (A.1) so (3) follows.

APPENDIX 2

Derivation of Linearisation Variance Estimator

Write $\hat{\theta} - \theta = \tau_1/(\tau_1 + \tau_2) - \tau_1/\tau_3$, where

$$\tau_1 = \sum_j I(f_j = 1), \tau_2 = \sum_j I(f_j = 2)\gamma_{1j}, \tau_3 = \sum_j F_j I(f_j = 1).$$

Let $\mu_t = E(\tau_t)$, $t = 1, 2, 3$, and note that $\mu_1 + \mu_2 = \mu_3$ from (3). A linearised expression for $\hat{\theta} - \theta$ is $\mu_1(-\tau_1 - \tau_2 + \tau_3)/\mu_3^2$, the variance of which may be expressed as

$$\begin{aligned}
 & \text{var}(\hat{\theta} - \theta) \\
 & \approx \text{var}\left[\left(\mu_1/\mu_3^2\right) \sum_{j=1}^J \{(F_j - 1)I(f_j = 1) - \gamma_{1j}I(f_j = 2)\}\right] \\
 &= (\mu_1/\mu_3^2)^2 \sum_{j=1}^J \left[(F_j - 1)^2 \text{Pr}(f_j = 1) + E\{\gamma_{1j}^2 I(f_j = 2)\}\right]. \quad (\text{A.2})
 \end{aligned}$$

This generalises the expression for the variance in Proposition 3 of SE. The expression for \hat{v} in (4) is obtained by replacing terms in (A.2) by their unbiased estimators.

First, μ_1 and μ_3 are estimated by τ_1 and $\tau_1 + \tau_2$ respectively so that μ_1/μ_3^2 is estimated by $\hat{\theta}/(\tau_1 + \tau_2)$. Next note that

$$\begin{aligned}
 & E\left[I(f_j = 3)(\gamma_{1j}^2 - \gamma_{2j})\right] \\
 &= \sum_{k=1}^{F_j} \sum_{\substack{\ell=1 \\ k \neq \ell \neq m}}^{F_j} \sum_{m=1}^{F_j} \pi_{(jk)} \pi_{(j\ell)} \pi_{(jm)} \left[\prod_{\substack{n=1 \\ n \neq k, \ell, m}}^{F_j} (1 - \pi_{(jn)}) \right] \beta_{(j\ell)} \beta_{(jm)} \\
 &= \sum_{k=1}^{F_j} \sum_{\substack{\ell=1 \\ k \neq \ell \neq m}}^{F_j} \sum_{m=1}^{F_j} \pi_{(jk)} \left[\prod_{\substack{n=1 \\ n \neq k}}^{F_j} (1 - \pi_{(jn)}) \right] \\
 &= (F_j - 1)(F_j - 2)\text{Pr}(f_j = 1),
 \end{aligned}$$

using the notation of Appendix 1. We may also show that

$$E[I(f_j = 2)\gamma_{1j}] = (F_j - 1)\text{Pr}(f_j = 1) \quad (\text{A.3})$$

by following the proof of (3) in Appendix 1, but omitting the summation over j . (Note that the sides of (3) are equal to the corresponding sides of (A.3) summed over j). Hence, an unbiased estimator of $(F_j - 1)^2 \text{Pr}(f_j = 1)$ is

$$I(f_j = 3)(\gamma_{1j}^2 - \gamma_{2j}) + I(f_j = 2)\gamma_{1j}.$$

It follows that the numerator of the expression for $\hat{v}/\hat{\theta}^2$ in (4) is unbiased for the second part of the expression on the right side of (A.2) (omitting $(\mu_1/\mu_3^2)^2$) as required.

REFERENCES

- BETHLEHEM, J.G., KELLER, W.J. and PANNEKOEK, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*. 85, 38-45.
- HINKINS, S., OH, H.L. and SCHEUREN, F. (1997). Inverse sampling design algorithms. *Survey Methodology*. 23, 11-21.
- SKINNER, C.J., and ELLIOT, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B*. 64, 855-867.
- SKINNER, C.J., and HOLMES, D.J. (1998). Estimating the re-identification risk per record for microdata. *Journal of Official Statistics*. 14, 361-372.
- WILLENBORG, L., and DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer.

Inference for Partially Synthetic, Public Use Microdata Sets

J.P. REITER¹

ABSTRACT

To avoid disclosures, one approach is to release partially synthetic, public use microdata sets. These comprise the units originally surveyed, but some collected values, for example sensitive values at high risk of disclosure or values of key identifiers, are replaced with multiple imputations. Although partially synthetic approaches are currently used to protect public use data, valid methods of inference have not been developed for them. This article presents such methods. They are based on the concepts of multiple imputation for missing data but use different rules for combining point and variance estimates. The combining rules also differ from those for fully synthetic data sets developed by Raghunathan, Reiter and Rubin (2003). The validity of these new rules is illustrated in simulation studies.

KEY WORDS: Confidentiality; Disclosure; Multiple imputation; Synthetic data.

1. INTRODUCTION

When releasing data to the public, statistical agencies seek to provide detailed data without disclosing respondents' sensitive information. To reduce the risk of disclosures, agencies typically alter the original data for public release, for example by recoding variables, swapping data, or adding random noise to data values (Willenborg and de Waal 2001). However, these methods can distort relationships among variables in the data set. They also complicate analyses for users: to analyze properly perturbed data, users should apply the likelihood-based methods described by Little (1993) or the measurement error models described by Fuller (1993). These are difficult to use for non-standard estimands and may require analysts to learn new statistical methods and specialized software programs.

An alternative approach was proposed by Rubin (1993): release fully synthetic data sets comprised entirely of multiply-imputed data rather than actual values. This can protect confidentiality, since identification of units and their sensitive data can be difficult when the released data are not actual, collected values. And, with appropriate imputation and estimation methods based on the concepts of multiple imputation (Rubin 1987), the approach can allow data users to obtain valid inferences using standard, complete-data statistical methods and software. Such inferences can be made using the methods developed by Raghunathan *et al.* (2003), whose rules for combining point and variance estimates differ from those of Rubin (1987). Other discussions and variants of synthetic data approaches appear in Little (1993); Fienberg, Steele and Makov (1996); Fienberg, Makov and Steele (1998); Dandekar, Cohen and Kirkendall (2002a); Dandekar, Domingo-Ferrer and Sebe (2002b); Franconi and Stander (2002, 2003); Poletti, Franconi and Stander (2002); Poletti (2003) and Reiter (2002, 2003).

Although no data producers have adopted the fully synthetic approach on a production basis yet, some have adopted a variant of the approach: release partially synthetic data sets comprising a mix of actual and multiply-imputed values. For example, to protect data in the U.S. Survey of Consumer Finances, the U.S. Federal Reserve Board replaces monetary values at high disclosure risk with multiple imputations, then releases a mixture of these imputed values and the unreplaced, collected values (Kennickel 1997). Another partially synthetic approach has been implemented by Abowd and Woodcock (2001) to protect data in longitudinal, linked data sets. They replace all values of some sensitive variables with multiple imputations, but leave other variables at their actual values. A third approach has been implemented by Liu and Little (2002), who develop an algorithm for simulating multiple values of key identifiers for selected units. All these partially synthetic approaches are appealing because they promise to maintain many of the benefits of fully synthetic data – protecting confidentiality while allowing users to make inferences without learning complicated statistical methods or software – with decreased sensitivity to the specification of imputation models.

Even though partially synthetic data sets are being publicly released, the literature does not contain technical results on how to obtain inferences from them. At first glance, it may appear appropriate to use the inferential methods for multiple imputation of missing data in Rubin (1987). Unfortunately, as shown in this article, these methods can result in biased variance estimates. Furthermore, and also as shown, the methods developed by Raghunathan *et al.* (2003) for analyzing fully synthetic data are not valid when applied on partially synthetic data. New methods of inference are required.

This paper describes methods for obtaining inferences from multiply-imputed, partially synthetic data sets. The derivation of these methods also provides prescriptions for

¹ J.P. Reiter, Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251. E-mail: jerry@stat.duke.edu

generating partially synthetic data. The paper is organized as follows. Section 2 presents the new methods of inference. Section 3 shows a derivation of these methods from a Bayesian perspective, and it discusses conditions under which the resulting inferences should be valid from a frequentist perspective. Section 4 describes simulation studies that illustrate the validity of these methods, as well as the ineffectiveness of competing rules for combining multiple point and variance estimates. Section 5 concludes with suggestions of future areas of research.

2. INFERENCES FROM MULTIPLY-IMPUTED, PARTIALLY SYNTHETIC DATA SETS

Let $I_j = 1$ if unit j is selected in the original survey, and $I_j = 0$ otherwise. Let $I = (I_1, \dots, I_N)$. Let Y_{obs} be the $n \times p$ matrix of collected (real) survey data for the units with $I_j = 1$; let Y_{nobs} be the $(N-n) \times p$ matrix of unobserved survey data for the units with $I_j = 0$; and, let $Y = (Y_{\text{obs}}, Y_{\text{nobs}})$. For simplicity, we assume that all sampled units fully respond to the survey. Let X be the $N \times d$ matrix of design variables for all N units in the population, e.g., stratum or cluster indicators or size measures. We assume that such design information is known approximately for all population units. It may come, for example, from census records or the sampling frame(s).

The agency releasing synthetic data, henceforth abbreviated as the imputer, constructs synthetic data sets based on the observed data, $D = (X, Y_{\text{obs}}, I)$, in a two-part process. First, the imputer selects the values from the observed data that will be replaced with imputations. Second, the imputer imputes new values to replace those selected values. Let $Z_j = 1$ if unit j is selected to have any of its observed data replaced with synthetic values, and let $Z_j = 0$ for those units with all data left unchanged. Let $Z = (Z_1, \dots, Z_N)$. Let $Y_{\text{rep},i}$ be all the imputed (replaced) values in the i -th synthetic data set, and let Y_{rep} be all unchanged (unreplaced) values of Y_{obs} . The $Y_{\text{rep},i}$ are assumed to be generated from the Bayesian posterior predictive distribution of $(Y_{\text{rep},i} | D, Z)$. The values in Y_{rep} are the same in all synthetic data sets. Each synthetic data set, d_i , then comprises $(X, Y_{\text{rep},i}, Y_{\text{rep}}, I, Z)$. Imputations are made independently for $i = 1, \dots, m$ times to yield m different synthetic data sets. These synthetic data sets are released to the public.

The values in Z can and frequently will depend on the values in D . For example, the imputer may choose to simulate sensitive variables or identifiers only for units in the sample with rare combinations of identifiers; or, the imputer may replace only those incomes above \$100,000 with imputed values. To avoid bias, imputers should account for such selections by imputing from the posterior predictive distribution of Y for those units with $Z_j = 1$. In practice, this can be done by using only the units with $Z_j = 1$ as the data when finding the posterior distributions for imputations.

Using all units with $I_j = 1$ can result in biased estimates or wider confidence intervals with overly conservative coverage rates, as illustrated in the simulations of section 4.

From these synthetic data sets, some user of the publicly released data, henceforth abbreviated as the analyst, seeks inferences about some estimand $Q = Q(X, Y)$, where the notation $Q(X, Y)$ means that Q is a function of (X, Y) . For example, Q could be the population mean of Y or the population regression coefficients of Y on X . In each synthetic data set d_i , the analyst estimates Q with some point estimator q and estimates the variance of q with some estimator v . It is assumed that the analyst determines the q and v as if the synthetic data were in fact collected data from a random sample of (X, Y) based on the actual survey design used to generate I .

For $i = 1, \dots, m$, let q_i and v_i be respectively the values of q and v in synthetic data set d_i . Under certain conditions to be described in section 3, the analyst can obtain valid inferences for scalar Q by combining the q_i and v_i . Specifically, the following quantities are needed for inferences:

$$\bar{q}_m = \sum_{i=1}^m q_i / m \quad (1)$$

$$b_m = \sum_{i=1}^m (q_i - \bar{q}_m)^2 / (m-1) \quad (2)$$

$$\bar{v}_m = \sum_{i=1}^m v_i / m. \quad (3)$$

The analyst then can use \bar{q}_m to estimate Q and

$$T_p = b_m / m + \bar{v}_m \quad (4)$$

to estimate the variance of \bar{q}_m . When q is a function of only (X, Y_{nrep}, I) and not any imputed values, the synthetic data inferences are identical to the observed data inferences; that is, the $q_i = q_{\text{obs}}$ and $v_i = v_{\text{obs}}$ for all i , and the $b_m = 0$. When n is large, inferences for scalar Q can be based on t -distributions with degrees of freedom $\nu_p = (m-1)(1+r_m^{-1})^2$, where $r_m = (m^{-1}b_m/\bar{v}_m)$. In many cases, r_m^{-1} and hence ν_p will be large enough that a normal distribution provides an adequate approximation to the t -distribution. Extensions for multivariate Q are not presented here.

T_p differs from the variance estimator for multiple imputation of missing data, $T_m = (1+1/m)b_m + \bar{v}_m$ (Rubin 1987). In the partially synthetic data context, the \bar{v}_m estimates $\text{Var}(q_{\text{obs}})$ and the b_m/m estimates the additional variance due to using a finite number of imputations. In the missing data context, the \bar{v}_m and b_m/m have the same interpretations, but an additional b_m is needed to average over the nonresponse mechanism (Rubin 1987, Chapter 4). This additional averaging is unnecessary in partially synthetic data settings, since the selection mechanism Z , which is set by the imputer, is not treated as stochastic.

T_p also differs from the variance estimator for analyzing fully synthetic data, $T_s = (1 + 1/m) b_m - \bar{v}_m$ (Raghunathan *et al.* 2003). To generate fully synthetic data, new units are sampled off the frame(s) for each synthetic data set, and their data are imputed. As shown by Raghunathan *et al.* (2003), this re-sampling and imputation process results in $b_m - \bar{v}_m$ as an appropriate estimate of $\text{Var}(q_{\text{obs}})$. For partially synthetic data, the original units are released for each data set, so that \bar{v}_m is an appropriate estimate of $\text{Var}(q_{\text{obs}})$.

3. JUSTIFICATION OF NEW COMBINING RULES

This section shows a Bayesian derivation of the inferences described in section 2 and conditions under which these inferences are valid from a frequentist perspective. These results are based on, and closely follow, the theory developed in Raghunathan *et al.* (2003).

3.1 Bayesian Derivation

For this derivation, we assume that the analyst and imputer use the same Bayesian model. The posterior distribution for $(Q|d^m)$, where $d^m = \{d_1, d_2, \dots, d_m\}$, can be decomposed as

$$f(Q|d^m) = \int f(Q|d^m, D, B) f(D|d^m, B) f(B|d^m) dD dB \quad (5)$$

where $B = \text{Var}(q_i|D, Z)$. The integration with respect to $f(D|d^m, B) dD$ is only over the values of Y_{obs} that are replaced with imputations; the (X, Y_{rep}, I) components of D remain fixed. Given D , the synthetic data are irrelevant, so that $f(Q|d^m, D, B) = f(Q|D)$. We assume standard Bayesian asymptotics hold, so that $f(Q|D) \sim N(q_{\text{obs}}, v_{\text{obs}})$, where q_{obs} and v_{obs} are the posterior mean and variance of Q determined using D .

Integrating (5) over D , we obtain $f(Q|d^m, B)$. Since only q_{obs} and v_{obs} are needed for inferences about $(Q|D)$, for $f(D|d^m, B)$ it is sufficient to determine $f(q_{\text{obs}}, v_{\text{obs}}|d^m, B)$. We assume imputations are made so that, for all i , $(q_i|D, B) \sim N(q_{\text{obs}}, B)$ and $(v_i|D, B) \sim (v_{\text{obs}}, << B)$. Here, the notation $F \sim (G, << H)$ means that the random variable F has a distribution with expectation of G and variability much less than H . In actuality, v_i is typically centered at a value larger than v_{obs} , since synthetic data incorporate uncertainty due to drawing values of the parameters. For large sample sizes n , this bias should be minimal. The assumption that $E(q_i|D, B) = q_{\text{obs}}$ should be reasonable when the imputations are drawn from the correct posterior distribution of Y for those units with $Z_j = 1$.

Assuming flat priors for q_{obs} and v_{obs} , standard Bayesian theory implies that $(q_{\text{obs}}|d^m, B) \sim N(\bar{q}_m, B/m)$ and $(v_{\text{obs}}|d^m, B) \sim (\bar{v}_m, << B/m)$. Hence, the posterior mean and variance of $(Q|d^m, B)$ are

$$\begin{aligned} E(Q|d^m, B) &= E(E(Q|D, d^m, B)|d^m, B) \\ &= E(q_{\text{obs}}|d^m, B) = \bar{q}_m \end{aligned} \quad (6)$$

$$\begin{aligned} \text{Var}(Q|d^m, B) &= E(\text{Var}(Q|D, d^m, B)|d^m, B) \\ &\quad + \text{Var}(E(Q|D, d^m, B)|d^m, B) \\ &= \bar{v}_m + B/m. \end{aligned} \quad (7)$$

Since all the convolutions involve normal distributions, $f(Q|d^m, B) \sim N(\bar{q}_m, \bar{v}_m + B/m)$.

To integrate this distribution over $f(B|d^m)$, we use the fact that $((m-1)b_m B^{-1}|d^m) \sim \chi_{m-1}^2$ and, following the approximation in Rubin (1987), fit the first two moments of $\bar{v}_m + B/m$ to a mean-square random variable. The resulting approximation to the posterior distribution of Q is $(Q|d^m) \sim t_{v_p}(\bar{q}_m, T_p)$, where v_p is as defined in section 2.

3.2 Randomization Validity

For inferences based on (1) - (4) to have valid frequentist properties, we require two conditions. First, the analyst must use randomization valid estimators, q and v . That is, when q and v are applied on D to get q_{obs} and v_{obs} , the $(q_{\text{obs}}|X, Y) \sim N(Q, U)$ and $(v_{\text{obs}}|X, Y) \sim (U, << U)$, where the relevant distribution is that of I . Second, the synthetic data generation methods must be proper in a sense similar to Rubin (1987). Specifically, the data generation methods should satisfy the following conditions:

C1: Averaging over imputations of $Y_{\text{rep}, i}$, it is required that

- (i) $(q_i|X, Y, I, Z) \sim N(q_{\text{obs}}, B)$;
- (ii) $(b_m|X, Y, I, Z) \sim (B, << B)$; and,
- (iii) $(\bar{v}_m|X, Y, I, Z) \sim (v_{\text{obs}}, << B/m)$, where $B = \text{Var}(q_i|X, Y, I, Z)$.

C2: Averaging over the sampling and replacement mechanisms $(I, Z|X, Y)$, it is required that $(B|X, Y) \sim (B_0, << U)$ where $B_0 = E(b_m|X, Y)$.

Essentially, these conditions require the synthetic data be generated so that the q_i are unbiased for q_{obs} , the b_m is unbiased for B_0 , and the \bar{v}_m is unbiased for v_{obs} . Further discussion of proper imputation can be found in Rubin (1987).

Using these assumptions, it follows that

$$\begin{aligned} E(\bar{q}_m|X, Y) &= E(E(\bar{q}_m|X, Y, I, Z)|X, Y) \\ &= E(q_{\text{obs}}|X, Y) = Q \end{aligned} \quad (8)$$

$$\begin{aligned} \text{Var}(\bar{q}_m|X, Y) &= E(\text{Var}(\bar{q}_m|X, Y, I, Z)|X, Y) \\ &\quad + \text{Var}(E(\bar{q}_m|X, Y, I, Z)|X, Y) \\ &= E(B|X, Y)/m + \text{Var}(q_{\text{obs}}|X, Y) = B_0/m + U. \end{aligned} \quad (9)$$

Since $(q_{\text{obs}}|X, Y)$ and the $(q_i|X, Y, I, Z)$ are assumed to have normal distributions, it follows that $(\bar{q}_m|X, Y) \sim N(Q, B_0/m + U)$.

When C1 and C2 hold, T_p is an unbiased estimator of $B_0/m + U$. The t -approximation is justified using the method outlined in Rubin (1987). Specifically, the t -approximation follows since $((m-1)b_m B_0^{-1}|X, Y) \sim \chi_{m-1}^2$, and the degrees of freedom of a chi-squared random variable equals two times the square of its expectation over its variance.

4. SIMULATION STUDIES

This section illustrates the validity of these new combining rules, as well as the ineffectiveness of T_m and T_s as variance estimators, using simulation studies of partially synthetic strategies. Section 4.1 describes two studies in which the imputer generates synthetic data only for selected units. Section 4.2 describes a study in which the imputer generates synthetic data for all values of one survey variable, leaving the others at their observed values. For illustrations, the simulations use artificial data and correct posterior distributions for imputations. Of course, in real settings the correct imputation model typically is not known and must be estimated using the observed data and subject-matter expertise. For all simulations, the population sizes are considered infinite so that finite population correction factors are ignored.

4.1 Imputation for Selected Units

Imputers may decide to replace the observed values for some units in the collected data, then release a mixture of the imputed and observed values. This strategy is employed in two simplistic although illustrative simulations, the first involving a single variable and the second four variables.

4.1.1 Simulations Using a Single Variable

Each observed dataset, D , comprises $n = 100$ values drawn randomly from $Y \sim N(0, 10^2)$. Two different schemes are used to specify the units with $Z_j = 1$, so that two sets of partially synthetic data sets are generated for each D . The first scheme, labelled "Random", replaces Y for 20 units randomly sampled from D . The second scheme, labelled "Big Y", replaces Y only for units with $Y_j > 10$.

For each D , and for each scheme, there are $m = 5$ synthetic data sets $d_i = (Y_{\text{rep},i}, Y_{\text{rep}}, I, Z)$, for $i = 1, \dots, 5$. The $Y_{\text{rep},i}$ are generated by using a Bayesian bootstrap (Rubin 1987, pages 123-124), which draws values of Y from a donor pool of selected values of Y_{obs} . Let Y_{elig} be the $n_0 \times 1$ vector of values of Y_{obs} that make up the donor pool. Let $n_{\text{rep}} = \sum_{j=1}^{100} Z_j$. The Bayesian bootstrap proceeds as follows:

1. Draw $(n_0 - 1)$ uniform random numbers. Sort these numbers in ascending order. Label these ordered numbers as $a_0 = 0, a_1, a_2, \dots, a_{n_0-1}, a_{n_0} = 1$.
2. Draw n_{rep} uniform random numbers, $u_1, u_2, \dots, u_j, \dots, u_{n_{\text{rep}}}$. For each of these u , impute $Y_{\text{elig},j}$ when $a_{j-1} < u \leq a_j$.

This Bayesian bootstrap is not likely to be used to impute data in real settings, since data sets contain more than one variable. It is used here because it provides straightforward, proper imputations for this illustration.

As mentioned in section 2, the correct posterior predictive distribution is $f(Y|D, Z)$, not $f(Y|D)$. This implies that the donor pool, Y_{elig} , should equal the set $\{Y_j; Z_j = 1\}$. This set is labelled "SELECT." For comparisons, synthetic values also are imputed using the donor set $\{Y_j; I_j = 1\}$. This set is labelled "ALL". Imputations based on ALL donors do not meet condition C1 in section 3.2, since $E(q_i|X, Y, I, Z) = \left(\sum_{j=1}^{100-n_{\text{rep}}} y_{\text{rep},j} + n_{\text{rep}} \bar{y}_{\text{obs}} \right) / n \neq \bar{y}_{\text{obs}}$, whereas imputations based on SELECT donors are proper.

Table 1 summarizes the results from 5,000 runs of this simulation. For both the Random and Big Y schemes, the averages of the \bar{q}_5 based on the SELECT donors approximately equal the average of q_{obs} . In the Random scheme, the \bar{q}_5 based on ALL donors is also unbiased, because $E(\bar{y}_{\text{rep}}|X, Y, I) = q_{\text{obs}}$ when averaged over Z (which is in fact stochastic in this scheme). However, when using ALL donors in the Big Y scheme, \bar{q}_5 has a large, negative bias. This results because imputed values are not restricted to be greater than 10 when using ALL donors.

In both the Random and Big Y schemes, 94.5% of the 5,000 synthetic 95% confidence intervals based on T_p and the SELECT donors cover zero. This rate is identical to the 94.5% coverage rate for the confidence intervals based on the observed data ($q_{\text{obs}} \pm 1.96\sqrt{v_{\text{obs}}}$). The nominal rates are less than 95% due to simulation error. The 2-3% difference between the averages of the T_p and the $\text{Var}(\bar{q}_5)$ roughly equals the difference between the average v_{obs} and $\text{Var}(q_{\text{obs}})$. The usual multiple imputation variance estimator, T_m , tends to overestimate the $\text{Var}(\bar{q}_5)$, leading to overly conservative confidence interval coverage rates, showing that T_m is not the correct variance estimator when analyzing properly imputed, partially synthetic data.

When imputations are based on ALL donors – an improper imputation method – in the Random scheme, T_p is negatively biased, and only 92.6% of the synthetic 95% confidence intervals cover zero. Using T_m increases the coverage rate to 95%, suggesting that it is safer to use T_m instead of T_p when ALL units are used for imputations. The confidence intervals based on ALL and T_m are on average wider than those based on SELECT and T_p . This illustrates the advantage of conditioning on Z to obtain proper imputations, even when the scheme used to set the $Z_j = 1$ does not depend on the values of Y .

Table 1
Simulation Results when Imputing Single Variable

Scheme and Imputation Method	Avg. \bar{q}_5	Var \bar{q}_5	Avg. T_p	Avg. T_m	Coverage of 95% CIs	
					Using T_p	Using T_m
$Z_j = 1$ for 20 randomly selected units						
SELECT	0.024	1.097	1.067	1.420	94.5%	96.7%
ALL	0.020	1.233	1.044	1.281	92.6%	94.9%
$Z_j = 1$ for units with $Y_j > 10$						
SELECT	0.016	1.031	1.011	1.068	94.5%	95.0%
ALL	-2.383	0.796	0.736	0.921	20.7%	28.8%
Observed data results*	0.016	1.021	1.000		94.5%	

* The column labels do not apply for this row. The average of the $q_{\text{obs}} = 0.016$, the variance of the $q_{\text{obs}} = 1.021$, the average of the $v_{\text{obs}} = 1.000$, and 94.5% of the five thousand 95% observed-data confidence intervals cover zero.

Although not shown in Table 1, the variance estimator for fully synthetic data, T_s , is negative in every one of the 5,000 simulations for both schemes and both imputation methods. Clearly, although valid for fully synthetic data (Raghunathan *et al.* 2003), T_s is not generally appropriate for partially synthetic data.

4.1.2 Simulations Using Four Variables

Each observed dataset, D , comprises $n = 200$ values of four variables, (Y_1, Y_2, Y_3, Y_4) , generated as follows: $(y_1, y_2, y_3) \sim MVN(\mathbf{0}, \Sigma)$, where Σ has all variances equal to one and all covariances equal to 0.5; and, $(y_4 | y_1, y_2, y_3) \sim N(10y_1 + 7y_2 + 4y_3, 25^2)$. To fix ideas, the variable Y_1 can be considered a key identifier and Y_4 the sensitive variable. The plan is to simulate values of the sensitive Y_4 for all units with “unusual” values of the key identifier, defined as $Y_1 > 1$. Hence, Y_{rep} comprises sampled values of (Y_1, Y_2, Y_3) and values of Y_4 for those units with $Y_1 \leq 1$. Typically, around 30 units per observed data set have $Y_1 > 1$.

As before, we examine two schemes for determining the posterior predictive distribution for imputations. SELECT uses only the units with $Z_j = 1$ as the data for the posteriors, and ALL uses all observed units. Imputations under each scheme are made by (i) drawing values of the parameters of the regression of Y_4 on (Y_1, Y_2, Y_3) from their posterior distribution, which is estimated using either the SELECT or ALL units, and (ii) drawing values of Y_4 for units with $Z_j = 1$ using the drawn values of parameters. There are $m = 5$ synthetic data sets generated for each observed data set D .

The estimands of interest include β , the regression coefficient of Y_1 in the linear regression of Y_4 on (Y_1, Y_2, Y_3) ; α , the regression coefficient of Y_4 in the regression of Y_1 on (Y_2, Y_3, Y_4) ; and \bar{Y}_4 , the population average of Y_4 . For inferences about β and α , q is the usual ordinary least squares estimator and v its variance estimator. For inferences about \bar{Y}_4 , q is the sample average and v its standard error.

Table 2 summarizes results from 5,000 runs of this simulation. When imputations are based on the SELECT units, the averages of the \bar{q}_5 and T_p are within simulation errors of the averages of the q_{obs} and $\text{Var}(\bar{q}_5)$. Additionally, the coverage rates for the synthetic 95% confidence intervals are similar to the coverage rates for the observed data 95% confidence intervals. The T_m are substantially larger than the $\text{Var}(\bar{q}_5)$, resulting in coverage rates around 97%. Although not shown in Table 2, T_s is negative in all 5,000 simulation runs. Taken together, these results are consistent with the findings in section 4.1.1: when imputations are drawn from a posterior distribution that conditions on Z , point and interval estimates based on T_p are more accurate than those based on T_m and T_s .

Although imputations based on ALL units are not proper, it is informative to examine the performances of T_p and T_m for such imputations. Imputers might base imputations on all observed units for practical reasons, for example because the units with $Z_j = 1$ do not provide sufficient data to fit the imputation models. The results mirror those in section 4.1.1: the T_p underestimate the $\text{Var}(\bar{q}_5)$, leading to coverage rates around 94%, whereas using T_m increases coverage rates to around 96%, primarily due to the positive bias in T_m . This again suggests that, when imputers do in fact base imputations on all observed units even though only some $Z_j = 1$, analysts are safer using T_m as the variance estimator rather than T_p . Just as seen in section 4.1.1, the intervals based on ALL units are typically wider than those based on SELECT units, suggesting that, when possible, imputers are better off basing imputations only on the units with $Z_j = 1$.

4.2 Imputation of all Values of Y for one Variable

Each observed data set comprises $n = 200$ values of four variables generated as follows: $(y_1, y_2, y_3) \sim MVN(\mathbf{0}, \mathbf{I})$ where \mathbf{I} is the identity matrix; and, $(y_4 | y_1, y_2, y_3) \sim N(10y_1 + 10y_2 + 10y_3, 25^2)$. Hence, the $Y_{\text{rep}} = (Y_1, Y_2, Y_3)$. Values of Y_4 are imputed from the Bayesian posterior predictive distribution of $(Y_4 | Y_{\text{obs}})$, derived by fitting the

regression of Y_4 on (Y_1, Y_2, Y_3) . All units have $Z_j = 1$ and are used as data for the posterior distributions. The estimands are the same as those described in section 4.1.2.

Table 3 summarizes the results from 5,000 simulation runs using $m = 5$ partially synthetic data sets. For all estimands, the averages of the \bar{q}_5 are practically identical to those of the q_{obs} . Additionally, the estimated variances based on T_p are close to the actual variances of the \bar{q}_5 . The slight upward bias results because \bar{v}_m tends to overestimate v_{obs} , as explained in section 3.1. The T_m on average overestimate the $\text{Var}(\bar{q}_5)$ by factors of more than two, and the T_s severely underestimate the $\text{Var}(\bar{q}_5)$ for α and \bar{Y}_4 . These problems are not due to small m ; in simulations with large m they persist. Although errors of these magnitudes may not occur in other settings, the results in this simple setting again indicate that T_m and T_s are not appropriate in general for analyzing partially synthetic data, especially when synthesizing entire variables.

Imputers have incentive to release small numbers of synthetic data sets. Each additional data set requires extra storage, and more importantly, releasing too many data sets

might jeopardize confidentiality if intruders somehow combine the imputed values to learn about the actual values. Table 4 displays results of independent replications of 5,000 simulation runs using different values of m . Point estimates are unbiased for all three estimands and so are not displayed in the table. The 95% confidence interval coverage rates are close to 95% for all values of m greater than two. The inflations in the T_p are again due to positive biases in the \bar{v}_m .

Table 4 illustrates that, when imputing entire variables, substantial efficiency gains can be made by increasing m beyond five. The amount of efficiency gain depends on the magnitude of b_m . When b_m is small relative to \bar{v}_m , for example when imputing values only for a small number of selected units, efficiency gains from increasing m will not be large. For any partially synthetic strategy, imputers can compare gains in efficiency with potential tradeoffs in confidentiality by simulation studies of intruder behavior on different numbers of released synthetic data sets.

Table 2
Simulation Results when Imputing Y_4 for Units with $Y_1 > 1$

Type of Inference	Avg. \bar{q}_5	Var \bar{q}_5	Avg. T_p	Avg. T_m	Coverage of 95% CIs	
					Using T_p	Using T_m
Estimand is β						
SELECT	10.02	5.45	5.68	8.97	95.3%	98.2%
ALL	10.04	5.89	5.28	7.57	93.7%	96.9%
Observed data*	10.00	4.70			95.5%	
Estimand is α						
SELECT	9.25×10^{-3}	4.49×10^{-6}	4.76×10^{-6}	6.97×10^{-6}	95.4%	97.9%
ALL	9.59×10^{-3}	5.03×10^{-6}	4.75×10^{-6}	6.31×10^{-6}	94.1%	96.5%
Observed data*	9.66×10^{-3}	4.26×10^{-6}			95.4%	
Estimand is \bar{Y}_4						
SELECT	-1.45×10^{-2}	4.97	5.01	6.09	95.0%	96.6%
ALL	-1.24×10^{-3}	5.19	4.82	5.59	93.8%	95.4%
Observed data*	-2.34×10^{-3}	4.76			94.5%	

* The column labels do not apply for this row. These are the averages of the q_{obs} , the variance of the q_{obs} , and the percentage of 95% observed-data confidence intervals that cover their Q .

Table 3
Simulation Results when Imputing an Entire Variable

Estimand	Avg. q_{obs}	Avg. \bar{q}_5	Var q_{obs}	Var \bar{q}_5	Avg. T_p	Avg. T_m	Avg. T_s
β	9.95	9.94	3.19	4.46	4.54	11.10	4.63
α	0.0137	0.0135	6.12	7.69	7.94	17.30	5.17
\bar{Y}_4	0.00	0.00	4.55	5.83	6.00	12.30	2.87

Table 4
Sensitivity of Partially Synthetic Inferences to Value of m

Setting	$\text{Var } \bar{q}_m$	Avg. T_p	95% CI cov.
Inference for β			
$m = 2$	6.52	6.50	92.7
$m = 3$	5.38	5.38	94.4
$m = 4$	4.64	4.89	95.4
$m = 5$	4.46	4.54	95.1
$m = 10$	3.87	3.88	94.4
$m = 50$	3.30	3.37	95.1
Inference for α			
$m = 2$	10.62	10.89	93.4
$m = 3$	8.92	9.15	94.9
$m = 4$	8.41	8.45	94.9
$m = 5$	7.69	7.94	95.4
$m = 10$	6.99	7.02	94.8
$m = 50$	6.05	6.28	95.5
Inference for \bar{Y}_4			
$m = 2$	8.13	7.96	93.4
$m = 3$	6.51	6.86	95.5
$m = 4$	6.11	6.33	95.6
$m = 5$	5.83	6.00	95.3
$m = 10$	5.13	5.38	95.4
$m = 50$	4.66	4.87	95.5

Variances associated with α are multiplied by 10^6 .

5. CONCLUDING REMARKS

The simulations in this article illustrate that the usual rules for combining multiply-imputed data sets can result in positively biased variance estimates when applied on partially synthetic data. The new rules presented here appear to remedy this problem, thereby leading to more reliable inferences. Further research is needed to assess the performance of these new rules when using partially synthetic strategies for genuine data, for which the correct imputation models are unlikely to be known. Additionally, evaluations of the new rules are needed when the released data sets also contain multiple imputations of missing data, for example imputations for item nonresponse. As conjectured by a referee of this article, when significant fractions of imputations are for missing data, T_m may not perform so unfavorably relative to T_p .

The simulations and theory also suggest that, when possible, imputers should use only units with values selected for replacement as the data when estimating posterior predictive distributions for imputations. Further examination of this prescription when simulating more than one variable in genuine data sets would be valuable.

Lastly, this article does not examine the implications of various partially synthetic data strategies for protecting confidentiality, nor does it compare partially synthetic approaches to alternative techniques for disclosure control.

Such comparisons would help imputers determine whether partially synthetic approaches are appropriate for their public use microdata releases.

ACKNOWLEDGEMENTS

This work was supported by the United States Bureau of the Census through a contract with Datametrics Research. The author thanks Trivellore Raghunathan, Donald Rubin, and Laura Zayatz for providing statistical support and general motivation for this research, and two referees and an associate editor for their valuable comments and suggestions.

REFERENCES

- ABOWD, J.M., and WOODCOCK, S.D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz and J. Theeuves (Eds.). *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: North-Holland. 215-277.
- DANDEKAR, R.A., COHEN, M. and KIRKENDALL, N. (2002a). Sensitive micro data protection using Latin hypercube sampling technique. In J. Domingo-Ferrer (Ed.). *Inference Control in Statistical Databases*. Berlin: Springer-Verlag. 117-125.
- DANDEKAR, R.A., DOMINGO-FERRER, J. and SEBE, F. (2002b). LHS-based hybrid microdata versus rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer (Ed.). *Inference Control in Statistical Databases*. Berlin: Springer-Verlag. 153-162.
- FIENBERG, S.E., MAKOV, U.E. and STEELE, R.J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*. 14, 485-502.
- FIENBERG, S.E., STEELE, R.J. and MAKOV, U.E. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: Data swapping and log-linear models. In *Proceedings of Bureau of Census 1996 Annual Research Conference*. 87-105.
- FRANCONI, L., and STANDER, J. (2002). A model based method for disclosure limitation of business microdata. *The Statistician*. 51, 1-11.
- FRANCONI, L., and STANDER, J. (2003). Spatial and non-spatial model-based protection procedures for the release of business microdata. *Statistics and Computing*. Forthcoming.
- FULLER, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*. 9, 383-406.
- KENNICKELL, A.B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson (Eds.). *Record Linkage Techniques, 1997*. Washington, D.C.: National Academy Press. 248-267.
- LITTLE, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*. 9, 407-426.

- LIU, F., and LITTLE, R.J.A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. In *Proceedings of the Survey Research Methods Section*, American Statistical Association. 2133-2138.
- POLETTINI, S. (2003). Maximum entropy simulation for microdata protection. *Statistics and Computing*. Forthcoming.
- POLETTINI, S., FRANCONI, L. and STANDER, J. (2002). Model-based disclosure protection. In J. Domingo-Ferrer (Ed). *Inference Control in Statistical Databases*. Berlin: Springer-Verlag. 83-96.
- RAGHUNATHAN, T.E., REITER, J.P. and RUBIN, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*. 19, 1-16.
- REITER, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*. 18, 531-544.
- REITER, J.P. (2003). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Tech. Rep., Institute of Statistics and Decision Sciences, Duke University.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- RUBIN, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*. 9, 462-468.
- WILLENBORG, L., and DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

The High Entropy Variance of the Horvitz-Thompson Estimator

K.R.W. BREWER and MARTIN E. DONADIO¹

ABSTRACT

Using both purely design-based and model-assisted arguments, it is shown that, under conditions of high entropy, the variance of the Horvitz-Thompson (HT) estimator depends almost entirely on first order inclusion probabilities. Approximate expressions and estimators are derived for this "high entropy" variance of the HT estimator. Monte Carlo simulation studies are conducted to examine the statistical properties of the proposed variance estimators.

KEY WORDS: Horvitz-Thompson estimator; Model assisted survey sampling; Monte Carlo simulation; Variance estimation.

1. INTRODUCTION

Let U denote a finite population of N units labelled $i = 1, \dots, N$, and let Y_i denote the value for the i -th unit of a certain characteristic y . Consider the problem of estimating the population total $Y_* = \sum_{i=1}^N Y_i$. If a sample, s , of n units is drawn without replacement from U with first order inclusion probabilities $\pi_i, i \in U$, the Horvitz-Thompson (HT) (1952) estimator of the total is $\hat{Y}_{HT} = \sum_{i \in s} Y_i \pi_i^{-1}$. In this paper, we confine consideration to fixed size sampling designs. For this important special case, Sen (1953) and Yates and Grundy (1953) showed independently that \hat{Y}_{HT} has the variance

$$V(\hat{Y}_{HT}) = (1/2) \sum_{i \in U} \sum_{j(*i) \in U} (\pi_i \pi_j - \pi_{ij}) (Y_i \pi_i^{-1} - Y_j \pi_j^{-1})^2, \quad (1)$$

where π_{ij} is the second order or joint inclusion probability of the i -th and j -th population units together in the same sample. They therefore suggested the variance estimator

$$\hat{V}_{SYG}(\hat{Y}_{HT})$$

$$= (1/2) \sum_{i \in s} \sum_{j(*i) \in s} \pi_{ij}^{-1} (\pi_i \pi_j - \pi_{ij}) (Y_i \pi_i^{-1} - Y_j \pi_j^{-1})^2. \quad (2)$$

This is known to perform better than the variance estimator proposed by Horvitz and Thompson (1952) (the latter, however, usually being unbiased for random n), but the critical dependence of (2) on π_{ij} has proved problematical (Brewer 1999). If one or more of the $N(N-1)/2$ distinct values of π_{ij} are zero, the estimator (2) is biased downwards. And if any of them should be very small compared with their corresponding values of $\pi_i \pi_j$, (2) will be unstable (that is, it will itself be subject to high variance). In addition, the double sum feature of (2) is quite inconvenient, especially for large sample sizes. Not only are there many more π_{ij} 's than there are π_i 's; it is also frequently the case that the individual π_{ij} 's are problematic to evaluate. In view of these difficulties, the aim of this paper is to provide

alternative variance estimators, which do not depend on the π_{ij} 's and are simple to compute.

In the next section, a new expression for the design-variance of the HT estimator is presented. This new expression leads, under high entropy conditions, to the derivation of an approximate formula for $V(\hat{Y}_{HT})$, which is π_{ij} -free. In section 3, we check the usefulness of our approximate formulae using a model assisted approach. An estimator of our approximate variance is proposed in section 4; this variance estimator is expected to perform well under conditions of high entropy (meaning the absence of any detectable pattern or ordering in the selected sample units). Most sample selection schemes though, result in the selection of high entropy samples. With the aim of testing the usefulness of the variance estimator presented in section 4, some empirical studies were conducted. The main findings from these studies are reported in section 5. Some concluding remarks are provided in section 6.

2. SOME APPROXIMATE FORMULAE FOR THE DESIGN-VARIANCE OF THE HT ESTIMATOR

We begin this section by presenting an alternative formulation for the variance of the HT estimator, valid only when the sampling design is of fixed size. Before proceeding, we state the following relations, which will be useful later:

$$\sum_{j(*i) \in U} \pi_{ij} = (n-1)\pi_i, \quad i \in U \quad (3)$$

$$\sum_{j(*i) \in U} \pi_i \pi_j = (n-\pi_i)\pi_i, \quad i \in U \quad (4)$$

$$\sum_{i \in U} \sum_{j(*i) \in U} \pi_{ij} = n(n-1) \quad (5)$$

¹ Ken Brewer, School of Finance and Applied Statistics, Australian National University, ACT 0200, Australia. E-mail: Ken.Brewer@anu.edu.au and Martin E. Donadio, Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia. E-mail: M.Donadio@abs.gov.au.

$$\sum_{i \in U} \sum_{j(*i) \in U} \pi_i \pi_j = n^2 - \sum_{i \in U} \pi_i^2. \quad (6)$$

The alternative formulation is obtained as follows. We start with a trivial modification of (1),

$$\begin{aligned} V(\hat{Y}_{\text{HT}}) &= (1/2) \sum_{i \in U} \sum_{j(*i) \in U} (\pi_i \pi_j - \pi_{ij}) \left\{ (Y_i \pi_i^{-1} - Y_{\cdot} n^{-1}) \right. \\ &\quad \left. - (Y_j \pi_j^{-1} - Y_{\cdot} n^{-1}) \right\}^2 \\ &= (1/2) \sum_{i \in U} \sum_{j(*i) \in U} (\pi_i \pi_j - \pi_{ij}) \left\{ (Y_i \pi_i^{-1} - Y_{\cdot} n^{-1})^2 \right. \\ &\quad \left. + (Y_j \pi_j^{-1} - Y_{\cdot} n^{-1})^2 - 2(Y_i \pi_i^{-1} - Y_{\cdot} n^{-1})(Y_j \pi_j^{-1} - Y_{\cdot} n^{-1}) \right\}. \end{aligned}$$

Using the relations (3) and (4), the above equation may be shown to be identical to

$$\begin{aligned} V(\hat{Y}_{\text{HT}}) &= \sum_{i \in U} \pi_i (Y_i \pi_i^{-1} - Y_{\cdot} n^{-1})^2 - \sum_{i \in U} \pi_i^2 (Y_i \pi_i^{-1} - Y_{\cdot} n^{-1})^2 \\ &\quad - \sum_{i \in U} \sum_{j(*i) \in U} (\pi_i \pi_j - \pi_{ij}) (Y_i \pi_i^{-1} - Y_{\cdot} n^{-1})(Y_j \pi_j^{-1} - Y_{\cdot} n^{-1}). \quad (7) \end{aligned}$$

The first term in (7) is virtually the same as the variance of the corresponding Hansen-Hurwitz (1943) estimator of total for sampling at n draws with replacement, the probability of selecting unit i at each draw being $p_i = \pi_i/n, i \in U$. The second term can be viewed as a finite population correction. Consequently, these two terms together plausibly constitute a first approximation to the entire variance of the HT estimator and, importantly, neither of them depends on the π_{ij} 's.

The magnitude of the third term depends mostly on the sampling design $p(s)$. Thus, if $p(s)$ is such that $\pi_{ij} \approx \pi_i \pi_j$ for all $i \neq j \in U$, then we can expect a very small third term in (7) (compared with the other two). This condition seems to be satisfied by high entropy sampling designs. For example, in simple random sampling without replacement (*srswor*), which maximizes the entropy among all fixed sized designs (see Hájek 1981), the second order inclusion probabilities can be written as $\pi_{ij} = \pi_i \pi_j [N(n-1)/(n(N-1))]$. The factor $N(n-1)/(n(N-1))$ is less than 1, and tends to 1 for large population and sample sizes. For this design, the third term in (7) accounts for only $1/N$ of the entire variance of the HT estimator. Furthermore, for several probability proportional-to-size designs, such as rejective sampling (Hájek 1964) and randomized systematic πps sampling (Hartley and Rao 1962), the condition $\pi_{ij} \approx \pi_i \pi_j$ also holds, provided N and n are large enough.

There are some exceptions, however, in which the third term in (7) can be important. The most important of these

exceptions is systematic sampling from a population in which the units are arranged in a meaningful order prior to the selection. In such a case, a number of second order inclusion probabilities can even be equal to zero. This and other special cases need to be dealt with separately, and are not discussed further in this paper.

The rest of this section is devoted to deriving an approximation to $V(\hat{Y}_{\text{HT}})$ that uses first order inclusion probabilities only. We start by proposing a simple approximation to the π_{ij} of the form

$$\pi_{ij} \approx \tilde{\pi}_{ij} = \pi_i \pi_j (c_i + c_j)/2, \quad i \neq j \in U. \quad (8)$$

Three possible choices for $c_i, i \in U$, are then:

$$c_i = (n-1)/(n-\pi_i), \quad (9)$$

$$c_i = c = (n-1) / \left(n - n^{-1} \sum_{k \in U} \pi_k^2 \right) \text{ and} \quad (10)$$

$$c_i = (n-1) / \left(n - 2\pi_i + n^{-1} \sum_{k \in U} \pi_k^2 \right). \quad (11)$$

The first two choices of c_i are prompted by ratios of sums of π_{ij} to the corresponding sums of $\pi_i \pi_j$. Thus, on the one hand, formula (9) is obtained by comparing (3) with (4). On the other hand, formula (10) is suggested by the comparison of (5) and (6). Finally, formula (11) is based on the asymptotic expressions for π_{ij} obtained by Hartley and Rao (1962) and by Asok and Sukhatme (1976) for randomized systematic πps sampling and for Sampford's (1967) procedure respectively. To order $O(n^3 N^{-3})$, both these asymptotic expressions simplify to

$$\tilde{\pi}_{ij} = \pi_i \pi_j \{ (n-1)/n \} \left\{ 1 + n^{-1}(\pi_i + \pi_j) - n^{-2} \sum_{k \in U} \pi_k^2 \right\},$$

which in turn implies $c_i = \{(n-1)/n\} (1 - 2n^{-1}\pi_i - n^{-2} \sum_{k \in U} \pi_k^2)$. Under *srswor*, however, this choice of c_i does not yield the exact formula for the π_{ij} 's. For this reason, the slightly different expression given by (11) is used here, $(1 - 2n^{-1}\pi_i + n^{-2} \sum_{k \in U} \pi_k^2)$ being the first two terms in the Taylor expansion of the reciprocal of $(1 + 2n^{-1}\pi_i - n^{-2} \sum_{k \in U} \pi_k^2)$ and *vice versa*.

The next step consists of replacing the π_{ij} 's in the third term of (7) by the approximation (8). This replacement yields

$$\begin{aligned} & - \sum_{i \in U} \sum_{j(*i) \in U} (\pi_i \pi_j - \pi_{ij}) (Y_i \pi_i^{-1} - Y_{\cdot} n^{-1})(Y_j \pi_j^{-1} - Y_{\cdot} n^{-1}) \\ & \approx - \sum_{i \in U} \sum_{j(*i) \in U} \pi_i \pi_j [1 - (c_i + c_j)/2] \\ & \quad (Y_i \pi_i^{-1} - Y_{\cdot} n^{-1})(Y_j \pi_j^{-1} - Y_{\cdot} n^{-1}) \\ & = \sum_{i \in U} (1 - c_i) \pi_i^2 (Y_i \pi_i^{-1} - Y_{\cdot} n^{-1})^2, \end{aligned}$$

and thus the variance of the HT estimator may be approximated by

$$\begin{aligned} \tilde{V}(\hat{Y}_{\text{HT}}) &= \sum_{i \in U} [\pi_i - \pi_i^2 + (1 - c_i) \pi_i^2] (Y_i \pi_i^{-1} - Y_{\cdot} n^{-1})^2 \\ &= \sum_{i \in U} \pi_i (1 - c_i \pi_i) (Y_i \pi_i^{-1} - Y_{\cdot} n^{-1})^2. \end{aligned} \quad (12)$$

This approximate variance has a very simple form. It is also without error under *srswor* for all the three choices of c_i presented above.

3. A MODEL ASSISTED CHECK ON THE USEFULNESS OF THE APPROXIMATE VARIANCE FORMULAE

Consider the following ratio model as a possible description of the population being sampled:

$$\begin{aligned} \xi: Y_i &= \beta \pi_i + \varepsilon_i; E_{\xi} \varepsilon_i = 0; E_{\xi} \varepsilon_i^2 = \sigma_i^2; \\ E_{\xi} (\varepsilon_i \varepsilon_j) &= 0, i \neq j; i, j \in U. \end{aligned} \quad (13)$$

This is a shorthand model. It is intended to reflect the situation where the expected values of the Y_i are *intrinsically* proportional to the values X_i of an auxiliary variable x , and the inclusion probabilities π_i are *chosen* to be proportional to the X_i . It is of course impossible for the Y_i to be directly dependent on the inclusion probabilities as such, since those probabilities may be set quite arbitrarily by the person designing the sample.

The prediction or model expectation under ξ of the approximate variance expression (12) is

$$\begin{aligned} E_{\xi} \tilde{V}(\hat{Y}_{\text{HT}}) &= E_{\xi} \sum_{i \in U} \pi_i (1 - c_i \pi_i) (Y_i \pi_i^{-1} - Y_{\cdot} n^{-1})^2 \\ &= E_{\xi} \sum_{i \in U} \pi_i (1 - c_i \pi_i) (\varepsilon_i \pi_i^{-1} - \varepsilon_{\cdot} n^{-1})^2 \\ &= \sum_{i \in U} \sigma_i^2 \left\{ \pi_i^{-1} - n^{-1} - c_i (1 - 2n^{-1} \pi_i) - n^{-2} \sum_{k \in U} c_k \pi_k^2 \right\}, \end{aligned} \quad (14)$$

where $\varepsilon_{\cdot} = \sum_{i \in U} \varepsilon_i$. Ideally, expression (14) should be equal to $E_{\xi} V(\hat{Y}_{\text{HT}})$, namely $\sum_{i \in U} \sigma_i^2 (\pi_i^{-1} - 1)$ (Godambe 1955; Godambe and Joshi 1965). This condition leads to the implicit formula

$$c_i = \left(1 - n^{-1} - n^{-2} \sum_{k \in U} c_k \pi_k^2 \right) / (1 - 2n^{-1} \pi_i),$$

which can be solved for c_i iteratively, starting with the trial value $c_i^{(1)} = (n-1)/n$. To $O(N^{-1})$, this iterative solution is identical to (11). Alternatively, a closed expression can be derived by putting (14) equal to $\sum_{i \in U} \sigma_i^2 (\pi_i^{-1} - 1)$ and then requiring that $c_i = c$ for all $i \in U$, in which case we obtain

$$c = (1 - n^{-1}) \sum_{i \in U} \sigma_i^2 / \sum_{i \in U} \sigma_i^2 \left(1 - 2n^{-1} \pi_i - n^{-2} \sum_{k \in U} \pi_k^2 \right). \quad (15)$$

Under *srswor*, (15) becomes $c = N(n-1)/\{n(N-1)\}$, which yields the exact expression for $V(\hat{Y}_{\text{HT}})$. Even without *srswor*, replacing σ_i^2 by $\sigma^2 \pi_i$ in (15) returns (10) for c . It is reassuring that the purely design-based analysis and the model-assisted one produce results in such close agreement.

4. ESTIMATING THE DESIGN-VARIANCE OF THE HT ESTIMATOR

The aim of this section is to propose a plausible sample estimator for the approximate design-variance of the HT estimator given in (12). One such estimator is

$$\hat{\tilde{V}}(\hat{Y}_{\text{HT}}) = \sum_{i \in s} (c_i^{-1} - \pi_i) (Y_i \pi_i^{-1} - \hat{Y}_{\text{HT}} n^{-1})^2, \quad (16)$$

which is arrived at by replacing each population sum in (12) by the corresponding HT estimator, and adjusting by the factor c_i^{-1} . This estimator has some attractive properties: (i) For all three choices of c_i , it reduces to the standard variance estimator in the case of *srswor*; (ii) it is simple to compute, since no double sums are involved; and (iii) using Taylor linearization technique, it can be shown that (16) is approximately design-unbiased for (12).

A further attractive property of the estimator (16) is the following. When c_i is specified by (9), we have

$$c_i^{-1} - \pi_i = (n - \pi_i)/(n - 1) - \pi_i = \{n/(n - 1)\} (1 - \pi_i). \quad (17)$$

The factor $(1 - \pi_i)$ is easily interpretable as a finite population correction, while the factor $n/(n - 1)$ has an entirely different role, which can be explained as follows. It is easy to see that $\hat{\beta} = \hat{Y}_{\text{HT}} n^{-1}$ is a model unbiased estimator of β in model (13). Let us write $\hat{\sigma}_i^2 = (Y_i - \hat{\beta} \pi_i)^2$, for all i . Then $(Y_i \pi_i^{-1} - \hat{Y}_{\text{HT}} n^{-1})^2 = (Y_i - \hat{\beta} \pi_i)^2 \pi_i^{-2} = \hat{\sigma}_i^2 \pi_i^{-2}$, $i \in U$. It is not difficult to show that the factor $n/(n - 1)$ removes the (model) bias from $\sum_{i \in s} (Y_i \pi_i^{-1} - \hat{Y}_{\text{HT}} n^{-1})^2 = \sum_{i \in s} \hat{\sigma}_i^2 \pi_i^{-2}$ as an estimator of $\sum_{i \in s} \sigma_i^2 \pi_i^{-2}$.

The choice of (9) to specify the value of c_i also renders particularly simple the calculation both of the HT estimate itself and of its estimated variance; for substituting (17) into (16) and expanding that expression into individual terms we obtain:

$$\begin{aligned} \hat{\tilde{V}}(\hat{Y}_{\text{HT}}) &= \{n/(n - 1)\} \left\{ \sum_{i \in s} Y_i^2 \pi_i^{-2} - n^{-1} \hat{Y}_{\text{HT}}^2 \right. \\ &\quad \left. - \sum_{i \in s} Y_i^2 \pi_i^{-1} + 2n^{-1} \hat{Y}_{\text{HT}} \sum_{i \in s} Y_i - n^{-2} \hat{Y}_{\text{HT}}^2 \sum_{i \in s} \pi_i \right\}. \end{aligned}$$

This formula involves six expressions, namely n , \hat{Y}_{HT} , $\sum_{i \in s} Y_i^2 \pi_i^{-2}$, $\sum_{i \in s} Y_i \pi_i^{-1}$, $\sum_{i \in s} Y_i$, and $\sum_{i \in s} \pi_i$, which are the sample sums of 1 (unity), $Y_i \pi_i^{-1}$, $Y_i^2 \pi_i^{-2}$, $Y_i^2 \pi_i^{-1}$, Y_i , and π_i respectively. If these individual terms are cumulated over every sample unit, then \hat{Y}_{HT} and $\hat{V}(\hat{Y}_{\text{HT}})$ can be evaluated together, using only a single pass of the sample data.

Note that, if non-response is present, a first order correction for it may be obtained by conditioning the sample on the achieved sample size, which we may denote here by n' . That would involve replacing the original first order inclusion probabilities, π_i , by the “adjusted inclusion probabilities”, $\pi'_i = \pi_i n' / n$. (This terminology has been taken from Furnival, Gregoire and Grosenbaugh (1987), where the same type of adjustment was used in a different context). The summations over the achieved sample, s' , would then be n' , $\sum_{i \in s'} Y_i \pi'_i{}^{-1}$, $\sum_{i \in s'} Y_i^2 \pi'_i{}^{-2}$, $\sum_{i \in s'} Y_i^2 \pi'_i{}^{-1}$, $\sum_{i \in s'} Y_i$, and $\sum_{i \in s'} \pi'_i$ respectively.

Beyond the properties listed above, a further study of (16) is possible with the aid of the model ξ of (13). The most desirable expression for the ξ -expectation of an estimator of $V(\hat{Y}_{\text{HT}})$ is $\sum_{i \in s} \sigma_i^2 \pi_i^{-1} (\pi_i^{-1} - 1)$, because this in turn has design-expectation $\sum_{i \in U} \sigma_i^2 (\pi_i^{-1} - 1)$, which is the lower bound for the anticipated variance of any unbiased estimator (Godambe 1955; Godambe and Joshi 1965). For all the three definitions of c_i , the ξ -expectation of (16) differs from $\sum_{i \in s} \sigma_i^2 \pi_i^{-1} (\pi_i^{-1} - 1)$ by terms of order $O(Nn^{-1})$. Although these “unwanted” terms have opposite signs and therefore tend to cancel, they are not entirely negligible, being only $O(N^{-1})$ smaller than the variance itself.

In view of this, a new version of c_i , which retained the (design) properties (i)-(iii) for (16) and provided a closer expression to $\sum_{i \in s} \sigma_i^2 \pi_i^{-1} (\pi_i^{-1} - 1)$ for the ξ -expectation of (16), was desirable. These requirements are satisfied by a c_i defined as follows:

$$c_i = (n-1) / \left\{ n - (2n-1)(n-1)^{-1} \pi_i + (n-1)^{-1} \sum_{k \in U} \pi_k^2 \right\}, \quad (18)$$

for all $i \in U$. With this definition of c_i , the ξ -expectation of (16) still contains some “unwanted” terms, but they now consist only of a single term of order $O(Nn^{-2})$ – which is therefore smaller than $V(\hat{Y}_{\text{HT}})$ by a factor of order $O(N^{-1}n^{-1})$ – and other terms of smaller magnitude still.

5. SOME EMPIRICAL RESULTS

With the aim of evaluating the performance of the variance estimator proposed in section 4, some empirical studies were conducted. Three other variance estimators were also included in these studies: (i) the SYG estimator, given in (2); (ii) the variance estimator suggested by Hájek (1964, page 1520),

$$\hat{V}_{\text{HAJ}}(\hat{Y}_{\text{HT}}) = \{n/(n-1)\} \sum_{i \in s} (1 - \pi_i)(Y_i \pi_i^{-1} - A_s)^2, \quad (19)$$

where $A_s = \sum_{i \in s} a_i Y_i \pi_i^{-1}$, $a_i = (1 - \pi_i) / \sum_{k \in s} (1 - \pi_k)$; and (iii) a slight modification of (19) proposed by Deville (1999),

$$\hat{V}_{\text{DEV}}(\hat{Y}_{\text{HT}}) = \frac{1}{1 - \sum_{i \in s} a_i^2} \sum_{i \in s} (1 - \pi_i)(Y_i \pi_i^{-1} - A_s)^2. \quad (20)$$

It is worth mentioning that the estimator (19) was originally intended only for a particular high entropy design, namely rejective sampling, and not for all the high entropy ones. Later on, however, this estimator was proposed for its use with some other high entropy designs. For example, Rosén (1997) suggested the use of (19) in the context of Pareto sampling.

The inclusion of the estimators (2), (19) and (20) in our empirical studies deserves a brief explanation. The SYG variance estimator would usually be the preferred choice if the π_{ij} were known and were neither zero nor very small compared with the corresponding $\pi_i \pi_j$. Under these conditions, it would then be natural to ask: Is there a significant difference, in terms of performance, between (2) and the simpler estimator (16)? On the other hand, a comparison with (19) and (20) is of interest because these two estimators share with (16) the simplicity and π_{ij} -free features. Thus, they are “competitors” in the same class.

The performance of a variance estimator can be assessed in different ways; here we will focus on *bias* and *stability*. The main findings from our studies are reported in the remainder of this section. We will consider two cases separately, namely $n = 2$ and $n > 2$.

5.1 Case $n = 2$

With the aim of testing the variance estimators under different situations, nine small populations were used in this study, most of which were also included in the stability studies carried out by Rao and Bayless (1969). Table 1 summarizes the main features of each population, including the coefficients of variation (CV) of y and x , and the correlation coefficient, ρ , between y and x . Here, y is the variable for which total estimates are sought, and x is an auxiliary variable that may be used for sample selection. Note that N varies from 10 to 20, $\text{CV}(x)$ from 0.14 to 0.73, and ρ from 0.49 to 0.94. This provides a good mixture of populations with different characteristics.

The inclusion probabilities are chosen to be proportional to x , i.e., $\pi_i = 2X_i / X_s$, for all i . Two sampling designs are considered here, namely Brewer’s (1963) procedure (BREWER) and Tillé’s (1996) elimination procedure (TILLÉ). For both procedures, the π_{ij} are simple to compute and, for these nine populations, they are strictly positive (this condition is not always satisfied by TILLÉ). Moreover, since $n = 2$, for any sample $s = \{i, j\}$ we have $p(s) = \pi_{ij}$. Hence we can obtain the exact statistical properties of any given variance estimator \hat{V} .

To this end, let S denote the set of all possible samples of size $n = 2$ from a population U . The expectation of \hat{V} is then defined as

$$E(\hat{V}) = \sum_{s \in S} p(s) \hat{V}(s),$$

and its standard error (SE) as

$$SE(\hat{V}) = \left\{ \sum_{s \in S} p(s) [\hat{V}(s) - E(\hat{V})]^2 \right\}^{1/2}.$$

For each of the two sampling designs mentioned above, Table 2 displays the *relative bias* $RB(\hat{V}) = E(\hat{V})/V(\hat{Y}_{HT}) - 1$, expressed as a percentage, of the six π_{ij} -free variance estimators. The first two of these estimators need no explanation; the other four correspond to (16) coupled with (9), (10), (11), and (18) respectively. Since for $n = 2$ (only), \hat{V}_{DEV} and $\hat{V}_{16.9}$ are identical, they both appear in the same row. In order to simplify the reading of the table, the smallest RB (in absolute terms) in each population and sampling design has been highlighted.

The results in Table 2 prompt the following comments: (i) the performance of the π_{ij} -free variance estimators is reasonably good for all populations, with the possible exception of Population 4. An examination of the relationship between x and y for this population reveals the presence of some curvature, with larger cities growing at a higher rate. There is also an outlier – city 26 – for which the

number of people almost tripled in the 10-year period between 1920 and 1930. Another interesting case is given by Populations 5 and 6. These two populations have identical definitions, thus one would expect to obtain similar results for them. However, the RB figures for Population 5 are considerably worse than those for Population 6, specially for BREWER. The only noticeable difference between these two populations is that Population 5 contains an outlier (Farm 14 in the reference provided). It would appear then that the presence of outliers may result in some additional bias in these variance estimators. (ii) The estimator $\hat{V}_{16.18}$ seems to be the best of the class, performing remarkably well in all situations, and showing the smallest bias figures (in absolute values) in most cases; (iii) The estimator $\hat{V}_{16.10}$ tends to exhibit the largest bias figures.

Regarding stability, Table 3 reports the *coefficient of variation* $CV(\hat{V}) = SE(\hat{V})/E(\hat{V})$, expressed as a percentage, of all the seven variance estimators. It can be seen that the π_{ij} -free variance estimators tend to be more efficient (lower CVs) than \hat{V}_{SYG} , although the gains are small. Otherwise, there is little to choose from among these variance estimators, even though $\hat{V}_{16.10}$ is the best performer in all but the last population.

Table 1
Description of the Nine Small Populations

Pop.	Source	y	x	N	CV(y)	CV(x)	ρ
1	Cochran (1963, page 325)	No. of persons per block	No. of rooms per block	10	0.15	0.14	0.65
2	Yates (1981, page 150) Kraals 26-38	No. of persons absent	Total no. of persons	13	0.67	0.47	0.72
3	Rao (1963, page 207)	Corn acreage in 1960	Corn acreage in 1958	14	0.39	0.43	0.93
4	Cochran (1963, page 156) Cities 19-33	No. of people in 1930	No. of people in 1920	15	0.67	0.69	0.94
5	Sampford (1962, page 61) Even units	Oat acreage in 1957	Total acreage in 1947	17	0.61	0.71	0.80
6	Sampford (1962, page 61) Odd units	Oat acreage in 1957	Total acreage in 1947	18	0.75	0.73	0.91
7	Yates (1981, page 153)	Vol. of timber	Eye-estimated vol. of timber	20	0.51	0.48	0.49
8	Sukhatme (1954, page 279) Circles 1-20	Wheat acreage	No. of villages	20	0.63	0.50	0.59
9	Horvitz and Thompson (1952, page 682)	No. of households	Eye-estimated no. of households	20	0.44	0.40	0.87

Table 2
RB (%) of Variance Estimators for $n = 2$

		Pop1	Pop2	Pop3	Pop4	Pop5	Pop6	Pop7	Pop8	Pop9
B	\hat{V}_{HAJ}	-1.04	-2.97	-2.60	-6.05	-3.64	0.08	-0.81	-1.48	1.13
R	$\hat{V}_{DEV}, \hat{V}_{16.9}$	-0.98	-2.52	-2.29	-5.21	-3.00	0.54	-0.63	-1.33	1.24
E	$\hat{V}_{16.10}$	-1.37	-3.55	-3.21	-7.16	-4.31	0.82	-0.94	-1.89	1.80
W	$\hat{V}_{16.11}$	-0.59	-1.49	-1.37	-3.26	-1.69	0.26	-0.31	-0.76	0.68
.	$\hat{V}_{16.18}$	-0.20	-0.46	-0.46	-1.31	-0.38	-0.01	0.00	-0.19	0.13
T	\hat{V}_{HAJ}	-1.06	-4.40	-1.07	-5.90	-1.86	-0.41	0.32	-1.10	0.82
I	$\hat{V}_{DEV}, \hat{V}_{16.9}$	-1.00	-3.94	-0.75	-5.03	-1.19	0.07	0.51	-0.95	0.93
L	$\hat{V}_{16.10}$	-1.39	-4.91	-1.68	-6.91	-2.47	0.33	0.19	-1.50	1.48
L	$\hat{V}_{16.11}$	-0.62	-2.98	0.17	-3.14	0.09	-0.20	0.83	-0.39	0.38
É	$\hat{V}_{16.18}$	-0.23	-2.02	1.10	-1.25	1.37	-0.46	1.15	0.17	-0.17

Table 3
CV (%) of Variance Estimators for $n = 2$

		Pop1	Pop2	Pop3	Pop4	Pop5	Pop6	Pop7	Pop8	Pop9
B	\hat{V}_{SYG}	123	126	118	245	138	127	158	127	133
R	\hat{V}_{HAJ}	121	119	115	238	131	125	155	124	134
E	$\hat{V}_{\text{DEV}}, \hat{V}_{16.9}$	121	119	115	238	131	125	155	124	134
W	$\hat{V}_{16.10}$	120	117	114	236	128	124	153	123	135
E	$\hat{V}_{16.11}$	122	122	116	241	133	126	157	125	133
R	$\hat{V}_{16.18}$	122	125	117	243	136	127	158	126	133
T	\hat{V}_{SYG}	123	143	118	248	147	131	164	131	134
I	\hat{V}_{HAJ}	121	118	115	238	128	125	155	124	134
L	$\hat{V}_{\text{DEV}}, \hat{V}_{16.9}$	121	118	115	238	128	125	155	124	134
L	$\hat{V}_{16.10}$	121	116	114	235	125	124	154	123	135
É	$\hat{V}_{16.11}$	122	121	115	240	130	125	157	125	133
	$\hat{V}_{16.18}$	122	123	116	243	133	126	159	126	133

5.2 Case $n > 2$

In this section, we adopt a standard Monte Carlo simulation approach to examine the performance of the variance estimators. Two real populations are used in this study. The first one is a population of 220 blocks (BL220) taken from Appendix E in Kish (1965). The dataset contains two variables: Y_i = no. of dwellings occupied by renters in block i , and X_i = total no. of dwellings in block i . Some features of this population are: $\text{CV}(y) = 1.05$, $\text{CV}(x) = 0.85$, and $\rho = 0.97$.

The second population comprises 281 municipalities (MU281), and is given in Särndal, Swensson, and Wretman (1992). The role of the study variable, y , is played by RMT85, revenues from the 1985 municipal taxation, while P75, the municipality population in 1975, is used as a measure of size. The main characteristics of this population are: $\text{CV}(y) = 1.06$, $\text{CV}(x) = 0.96$, and $\rho = 0.99$.

Samples of sizes $n = 10, 20$ and 40 with $\pi_i \propto X_i$, $i \in U$, are drawn from BL220 and MU281 by means of randomized systematic πps sampling (RANSYS) and TILLÉ. For each sample, we compute a total estimate using the HT estimator, and variance estimates using the seven variance estimators mentioned in the previous section (for RANSYS, however, the Hartley and Rao (1962) approximation to the π_{ij} instead of the exact π_{ij} is used in formula (2)). This sampling-estimation process is repeated $R=50,000$ times.

Table 4 shows the observed Monte Carlo relative biases of the variance estimators for RANSYS and TILLÉ. Note that, for TILLÉ, no values have been provided in the row corresponding to the SYG variance estimator. This is because, given the populations, measures of size, and sample sizes employed here, TILLÉ produces strictly positive π_{ij} , which means that the SYG variance estimator is design unbiased. All the figures in this table are reasonably small, which seems to support our belief that, under conditions of high entropy, the calculation of the π_{ij} is not essential for obtaining nearly unbiased variance estimates.

Within the group of π_{ij} -free estimators, there are no noticeable differences among them so far as RANSYS is concerned, but \hat{V}_{HAJ} and its relative, \hat{V}_{DEV} , seem to perform somewhat better than the $\hat{V}_{16.*}$ family so far as TILLÉ is concerned, especially for $n = 40$. However, all the observed TILLÉ biases are positive and tend to increase as the sample size increases. This seems to indicate that TILLÉ is slightly lower in entropy than RANSYS, in which case the higher observed biases for the $\hat{V}_{16.*}$ family are reflecting the actual facts quite accurately.

Table 4
RB (%) of Variance Estimators for $n > 2$

Variance estimators	RANSYS			TILLÉ		
	$n = 10$	$n = 20$	$n = 40$	$n = 10$	$n = 20$	$n = 40$
BL220						
\hat{V}_{SYG}	0.13	1.02	-0.27	—	—	—
\hat{V}_{HAJ}	-0.14	0.47	-2.35	1.49	2.18	3.27
\hat{V}_{DEV}	-0.12	0.54	-2.15	1.52	2.25	3.48
$\hat{V}_{16.9}$	-0.06	0.83	-0.52	1.58	2.54	5.21
$\hat{V}_{16.10}$	-0.23	0.64	-0.75	1.41	2.34	4.97
$\hat{V}_{16.11}$	0.11	1.02	-0.30	1.75	2.73	5.45
$\hat{V}_{16.18}$	0.13	1.03	-0.29	1.77	2.74	5.45
MU281						
\hat{V}_{SYG}	-0.27	-0.43	0.77	—	—	—
\hat{V}_{HAJ}	-0.40	-0.75	-0.59	0.64	1.01	1.93
\hat{V}_{DEV}	-0.37	-0.68	-0.39	0.67	1.09	2.14
$\hat{V}_{16.9}$	-0.34	-0.51	0.67	0.70	1.26	3.22
$\hat{V}_{16.10}$	-0.40	-0.58	0.58	0.63	1.19	3.13
$\hat{V}_{16.11}$	-0.27	-0.43	0.76	0.77	1.34	3.31
$\hat{V}_{16.18}$	-0.27	-0.43	0.76	0.78	1.34	3.32

In order to test whether TILLÉ is of slightly lower entropy than RANSYS or not, we compared their Monte

Carlo variances (MCV) with formula (12), the high entropy approximation to the HT variance. The most accurate version of c_i , that is (18), was used to compute (12). The comparison is presented in Table 5. It is seen that the TILLE variances are somewhat smaller than the corresponding RANSYS variances. Moreover, the approximate variances provided by (12) are in closer agreement with the RANSYS variances. These findings support our previous conjecture that the entropy for TILLE is slightly lower than that for RANSYS, particularly when the finite population correction is appreciable.

Table 5
Comparison of Variances (all values in 10^4)

	BL220			MU281		
	$n = 10$	$n = 20$	$n = 40$	$n = 10$	$n = 20$	$n = 40$
(12)+(18)	14.06	6.572	2.830	565.5	264.3	113.7
MCV-RANSYS	14.07	6.520	2.841	566.2	265.3	112.8
MCV-TILLE	13.87	6.404	2.691	560.0	257.6	108.9

Next we focus on stability. Table 6 reports the observed Monte Carlo SE of the variance estimators. Clearly, there are no differences worth mentioning among the variance estimators. The same is true for a comparison of the two sampling procedures. It seems that stability does not constitute a relevant factor when choosing between these variance estimators.

Table 6
CV (%) of Variance Estimators for $n > 2$

Variance estimators	RANSYS			TILLE		
	$n = 10$	$n = 20$	$n = 40$	$n = 10$	$n = 20$	$n = 40$
BL220						
\hat{V}_{SYG}	58.31	41.16	30.70	57.43	40.41	29.54
\hat{V}_{HAJ}	57.90	40.49	29.48	57.39	40.24	29.08
\hat{V}_{DEV}	57.90	40.49	29.48	57.39	40.24	29.08
$\hat{V}_{16.9}$	57.02	40.54	29.64	57.41	40.29	29.24
$\hat{V}_{16.10}$	57.79	40.45	29.56	57.29	40.19	29.16
$\hat{V}_{16.11}$	58.04	40.64	29.73	57.53	40.39	29.32
$\hat{V}_{16.18}$	58.05	40.65	29.73	57.55	40.39	29.32
MU281						
\hat{V}_{SYG}	54.90	37.29	25.33	55.07	37.50	25.45
\hat{V}_{HAJ}	54.69	36.98	24.96	54.79	37.07	24.78
\hat{V}_{DEV}	54.68	36.98	24.95	54.79	37.07	24.77
$\hat{V}_{16.9}$	54.67	36.92	24.70	54.77	37.01	24.52
$\hat{V}_{16.10}$	54.63	36.89	24.66	54.74	36.98	24.48
$\hat{V}_{16.11}$	54.70	36.95	24.74	54.81	37.04	24.56
$\hat{V}_{16.18}$	54.71	36.96	24.74	54.81	37.04	24.56

6. SUMMARY

Estimators are derived for what, in the context of any high entropy selection procedure, is a close approximation to the design variance of the HT estimator of a total.

These estimators resemble, but are not identical to other variance estimators suggested for certain particular high entropy selection procedures by Hájek (1964), Rosen (1997), and Deville (1999). All these estimators have the important advantage over the standard SYG variance estimator that their formulae do not involve the second order inclusion probabilities, π_{ij} .

Empirical investigations indicate that these estimators all behave acceptably well, both for the important special case $n = 2$ and when n takes larger values. The estimator given by (16) with c_i defined by (18), which has certain near-optimal theoretical properties, appears to be noticeably less biased than the others for $n = 2$, but not for larger values of n .

For the case $n > 2$, two high entropy procedures were used, namely systematic sampling from a randomly ordered population (RANSYS) and the procedure proposed by Tillé (1996) (TILLÉ). The biases in all the variance estimators were consistently higher (meaning algebraically larger) for TILLÉ than for RANSYS, and particularly so when n took its largest value of 40. The differences between the TILLÉ biases and the RANSYS biases were also positive for all values of n , and again particularly so when $n = 40$. We conjecture that these differences may indicate that TILLÉ is a slightly lower entropy (and typically lower variance) selection procedure than RANSYS.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. P.S. Kott for suggesting equation (10) in a private communication, and an anonymous referee for three other suggestions that have added value to this paper.

REFERENCES

ASOK, C., and SUKHATME, B.V. (1976). On sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association*.71, 912-918.

BREWER, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*. 5, 5-13.

BREWER, K.R.W. (1999). Cosmetic calibration for unequal probability samples. *Survey Methodology*. 25, 205-212.

COCHRAN, W.G. (1963). *Sampling Techniques*. 2nd Ed. New York: John Wiley & Sons, Inc.

DEVILLE, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*. 25, 193-203.

- FURNIVAL, G.M., GREGOIRE, T.G. and GROSENBAUGH, L.R. (1987). Adjusted inclusion probabilities with 3P sampling. *Forest Science*. 33, 617-631.
- GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society B*. 17, 269-278.
- GODAMBE, V.P., and JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations I, II, and III. *Annals of Mathematical Statistics*. 36, 1707-1742.
- HÁJEK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*. 35, 1491-1523.
- HÁJEK, J. (1981). *Sampling from a finite population*. New York: Marcel Dekker.
- HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*. 14, 333-362.
- HARTLEY, H.O., and RAO, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*. 33, 350-374.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 47, 663-685.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- RAO, J.N.K. (1963). On three procedures of unequal probability sampling without replacement. *Journal of the American Statistical Association*. 58, 202-215.
- RAO, J.N.K., and BAYLESS, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *Journal of the American Statistical Association*. 64, 540-559.
- ROSÉN, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*. 62, 159-191.
- SAMPFORD, M.R. (1962). *An Introduction to Sampling Theory*. Edinburgh and London: Oliver and Boyd Ltd.
- SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*. 54, 499-513.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SEN, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*. 5, 119-127.
- SUKHATME, P.V. (1954). *Sampling Theory of Surveys with Applications*. Ames, Iowa: Iowa State College Press.
- TILLÉ, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika*. 83, 238-241.
- YATES, F. (1981). *Sampling Methods for Censuses and Surveys*. 4th Ed. London: Charles Griffin and Co.
- YATES, F., and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*. 15, 235-261.

Estimation with Link – Tracing Sampling Designs – A Bayesian Approach

MOSUK CHOW and STEVEN K. THOMPSON¹

ABSTRACT

In link-tracing designs, social links are followed from one respondent to another to obtain the sample. For hidden and hard-to-access human populations, such sampling designs are often the only practical way to obtain a sample large enough for an effective study. In this paper, we propose a Bayesian approach for the estimation problem. For studies using link-tracing designs, prior information may be available on the characteristics that one wants to estimate. Using this information effectively via a Bayesian approach should yield better estimators. When the available information is vague, one can use noninformative priors and conduct a sensitivity analysis. In our example we found that the estimators were not sensitive to the specified priors. It is important to note that, under the Bayesian setup, obtaining interval estimates to assess the accuracy of the estimators can be done without much added difficulty. By contrast, such tasks are difficult to perform using the classical approach. In general, a Bayesian analysis yields one distribution (the posterior distribution) for the unknown parameters, and from this a vast number of questions can be answered simultaneously.

KEY WORDS: Link-tracing designs; Snowball samples; Adaptive sampling; Graph sampling; Network sampling; Beta prior.

1. INTRODUCTION

Social network data include measurements on the relationships between people or other social entities as well as measurements on entities themselves. Collecting network data on entire networks requires a great deal of time and effort, especially when networks are large. It is thus important to be able to estimate network properties from samples. In link-tracing sampling designs, social links are followed from one respondent to another to obtain the sample. For hidden and hard-to-access human populations, such sampling designs are often the only practical way to obtain a sample large enough for an effective study. For example, in a study of injection drug use in relation to the spread of the HIV infection, social leads from initial respondents may be traced and the linked individuals added to the sample. (*e.g.*, see Neaigus, Friedman, Goldstein, Ildefonso, Curtis and Jose 1995; Neaigus, Friedman, Jose, Goldstein, Curtis, Ildefonso and Des Jarlais 1996 and Thompson and Collins 2002). Similarly, for studies of homeless people, respondents may be asked about other homeless people who will then be sampled.

Populations with social structure are often modeled as graphs, with the nodes of the graph representing populations and the arcs of the graph representing social links, relationships, or transactions. In the graph setting, the variables of interest include both those associated with nodes and those associated with pairs of nodes. The population graph itself can be viewed either as a fixed structure or as a realization of a stochastic graph model. Samples are taken to obtain information about the population graph. Usually, the sampling method will take advantage of the arcs or links from one entity to another.

There is a large literature on network sampling, both applied and theoretical. Frank (1977a, 1977b, 1977c, 1978, 1979, 1980, 1997) has many important results in sampling for social networks. His classic work (Frank 1971) presents basic solutions for estimating graph quantities from the sample data. Snijders and Nowicki (1997) propose various statistical approaches, including a Bayesian approach, for estimation and prediction with stochastic blockmodels for graphs in which the node values are not observed.

Snowball sampling (Goodman 1961) is one type of link-tracing sampling design in which individuals in an initial sample are asked to identify acquaintances, who in turn were asked to identify acquaintances, and so on for a fixed number of stages or waves. Erickson (1978) and Frank (1979) review snowball sampling designs with the goal of understanding how other “chain methods” (methods designed to trace ties through a network from a source to an end) can be used in practice. Snijders (1992) used the same term “snowball sampling” to include designs in which only a subsample of links from each node is traced. Frank and Snijders (1994) consider model and design-based estimation of a hidden population size, that is, the number of nodes in the graph, based on snowball samples. Another link-tracing procedure for which design-based estimators are available is adaptive cluster sampling (Thompson and Seber 1996), which has been formulated in the graph setting as well as the spatial setting.

With a fixed-population, design-based approach in the graph setting, both the characteristics of the people and the social network structure of the population are viewed as fixed, unknown values. The properties such as design-unbiasedness do not depend on any assumptions about the

¹ Mosuk Chow and Steven K. Thompson, Department of Statistics, The Pennsylvania State University, 326 Thomas Building, University Park, PA 16802, U.S.A.

population itself but they do depend on the sampling design being carried out as specified. In this paper, we consider the model-based methods since they can be applied to a wide range of sample selection procedures. In many studies of hidden and hard-to-reach populations, the sample selection procedures, including link-tracing, are not readily analyzed based on idealized design induced probabilities, but results from the model-based methods can be applied for the cases.

Thompson and Frank (2000) used a model-based approach to inference with link-tracing designs. In their paper, maximum likelihood estimators of population graph parameters and predictors of realized population graph quantities were described. In this paper, we adopt a Bayesian approach for the graph estimation problem. For real problems with sampling designs that follow social links from one person to another, prior information may be available on the characteristics that one wants to estimate. Using this information effectively via a Bayesian approach should yield improved estimators. Moreover, when the available information is vague, we can use noninformative priors and conduct a sensitivity analysis. It is important to note that under the Bayesian setup, obtaining interval estimates to assess the accuracy of the estimators can be done without much added difficulty whereas such tasks would be difficult to perform using the maximum likelihood approach. We deal with inferences for both the characteristic of nodes and also of arcs such as the prevalence of disease in a certain community and also the transmission rate of that disease between two subjects.

Notation for a full graph model with links related to node values and its likelihood function will be given in section 2. In section 3, the likelihood function for the sample obtained from a link-tracing design will be presented and a Bayesian inference method will be introduced. In section 4, an illustrative example will be given. The paper will be concluded by an empirical example and a discussion in section 5.

2. THE MODEL

Using notation similar to Frank (1971) and Thompson and Frank (2000), we denote the full set of node labels by $U = \{1, 2, \dots, N\}$ which form the population of N units. A variable of interest associated with an individual node u will be denoted Y_u while a variable of interest associated with pair of nodes u and v will be denoted A_{uv} . The sequence of node variables of interest is denoted by $\mathbf{Y} = (Y_1, \dots, Y_N)$. Here we consider the variable of interest A_{uv} as an indicator variable which equals one if there is an arc (directional link) from u to v and zero otherwise for two distinct nodes u and v . The matrix of arc indicators, having A_{uv} as the element in the u -th row and v -th column, is the graph adjacency matrix, denoted \mathbf{A} . For convenience we will assume that the diagonal elements A_{uu} are zero. The ordered pair (u, v) is referred to as a dyad of type

$(Y_u, Y_v; A_{uv}, A_{vu})$. In the following assumed model the node variables Y_1, \dots, Y_N are independent, identically distributed (i.i.d.) Bernoulli random variables with probabilities $P(Y_u = i) = \theta_i$, for $i = 0, 1$, and $\theta_0 + \theta_1 = 1$. Conditional on the node values Y_1, \dots, Y_N , the dyads (A_{uv}, A_{vu}) are independent, for $1 \leq u < v \leq N$, with conditional distribution given by $P[(A_{uv}, A_{vu}) = (k, l) | Y_u = i, Y_v = j] = \lambda_{ijkl}$ for all combinations of $i = 0, 1$; $j = 0, 1$; $k = 0, 1$; and $l = 0, 1$. For all combinations of i and j , the sums over k and l are denoted $\lambda_{ij\cdot\cdot} = \sum_k \sum_l \lambda_{ijkl}$ and equal 1. In order to get graph probabilities not depending on node identities, the following natural symmetry conditions are assumed: $\lambda_{1110} = \lambda_{1101}$, $\lambda_{1011} = \lambda_{0111}$, $\lambda_{1010} = \lambda_{0101}$, $\lambda_{1001} = \lambda_{0110}$, $\lambda_{0010} = \lambda_{0001}$ and $\lambda_{1000} = \lambda_{0100}$. For example, the first and the fifth conditions say that between two nodes having the same value, the probability of an arc in either direction is the same. Let N_i denote the total number of nodes with value i in the graph so that $N_0 + N_1 = N$. Let further M_{ijkl} denote the total number of dyads of type $(ijkl)$, that is, the total number of ordered node pairs (u, v) such that $(Y_u, Y_v; A_{uv}, A_{vu}) = (ijkl)$. The likelihood for the full graph under the model with parameters (θ, λ) is $L(\theta, \lambda; Y, A) = (\prod_{i=0}^1 \theta_i^{N_i}) (\prod_{i=0}^1 \prod_{j=0}^1 \prod_{k=0}^1 \prod_{l=0}^1 \lambda_{ijkl}^{M_{ijkl}})$.

3. BAYESIAN INFERENCE FROM LINK-TRACING DESIGNS

3.1 Likelihood Function given the Sample Data

A sample s from the graph is a subset of nodes from U and a subset of node pairs from U^2 . The sample data $d = (s, y_s, a_s)$ are a function of the sample selected and of the graph values y and a . For any design in which the selection of the sample depends on graph y and a values only through those values y_s and a_s included in the data, the design does not affect the value of estimators or predictors based on direct likelihood methods such as maximum likelihood or Bayes estimators (Rubin 1976, Thompson and Frank 2000). For example, many of the snowball and other link-tracing designs are ignorable for likelihood-based inference provided the selection procedure for the initial sample is ignorable. Any carefully implemented conventional or adaptive survey design would be ignorable in this sense. Nonignorable initial samples can occur when the selection is uncontrolled and selection probabilities are related to unobserved node and link values, as when people with risk-averse behaviors and low numbers of relationships are less conspicuous to investigators, thereby influencing what units are missed and hence influencing sample selection probabilities in ways that are not measured.

Consider the link-tracing design in which an initial sample s_0 is selected and all links out from nodes in s_0 are followed to add the set s_1 of nodes not in s_0 that are adjacent to nodes in s_0 . The whole sample is $s = s_0 \cup s_1$. The entire set of labels in the population can be written as

the union of three disjoint sets, $U = s_0 \cup s_1 \cup \bar{s}$ where \bar{s} denotes the nonsampled nodes. Here, we consider a design in which the decision to follow the links from node u depends on the node value y_u . For example, in a study on injection drug use, the initial sample may contain both users and nonusers. If the investigators choose to follow social links only from users, then the design depends adaptively on the node y -values as well as the links. The design then can be written $P(s | y_s, a_{s_{0U}})$, since the selection procedure depends on both node and link values. The data are $d = (s, y_s, a_{s_{0U}})$. Since the decision depends on y and a values only through the observed data, the design factors out of the likelihood function and divides out of the Bayes posterior, so that likelihood or Bayes inference depends only on the assumed model.

With the graph model described in the previous section, it then follows (Thompson and Frank 2000) that the likelihood with the sample data is:

$$L(\theta, \lambda; d) = P(s | y_s, a_{s_{0U}}) \sum \left(\prod_{u=1}^N \theta_{y_u} \right) \left(\prod_{u < v} \lambda_{y_u y_v a_{uv}, a_{vu}} \right)$$

where the sum is over all values of y_u and a_{uv} that are not fixed by the sample data.

For link-tracing designs in which all links, rather than a subsample, from the initial sample nodes are traced, all of the elements in the submatrix $a_{s_0 \bar{s}}$ are zero. It has been shown by Thompson and Frank (2000) that the likelihood function can then be written as:

$$L(\theta, \lambda; Y, A) = P(s | y_s, a_{s_{0U}}) \left(\prod_i \theta_i^{n_i(s)} \right) \left(\prod_{ijkl} \lambda_{ijkl}^{m_{ijkl}(s_0, s_0)} \right) \left(\prod_{ijk} \lambda_{ijk}^{m_{ijk}(s_0, s_1)} \right) \times \left[\sum_j \theta_j \prod_i \lambda_{ij0}^{n_i(s_0)} \right]^{n(\bar{s})} \quad (1)$$

where $n_i(s)$, $n_i(s_0)$, and $n_i(\bar{s})$ denote the numbers of nodes of type i in the full sample s , the initial sample s_0 , and the nonsampled nodes \bar{s} , respectively, and $m_{ijkl}(s_0, s_0)$, $m_{ijkl}(s_0, s_1)$ are the counts of node pairs in $s_0 \times s_0$ and $s_0 \times s_1$.

For a symmetric model, $\lambda_{ijkl} = 0$ for $k \neq l$ so that arcs are always two-way or, equivalently, they can be considered as undirected edges. The full symmetric model has parameters $\lambda_{ijkk} = \lambda_{jikk}$ for $i, j, k = 0, 1$, with $\lambda_{ij00} + \lambda_{ij11} = 1$. To simplify notation for this model, let $\beta_{i+j} = \lambda_{ij11}$ and thus β_k denotes the probability of a mutual link between two nodes having total value k , for $k = 0, 1$ or 2 . The above likelihood simplifies to

$$L(\theta, \beta; d) = P(s | y_s, a_{s_{0U}}) \left(\prod_i \theta_i^{n_i(s)} \right) \left(\prod_{i,j} \beta_{i+j}^{m_{ij00}(s_0, s)} (1 - \beta_{i+j})^{m_{ij00}(s_0, s)} \right) \times \left[\sum_j \theta_j \prod_i (1 - \beta_{i+j})^{n_i(s_0)} \right]^{n(\bar{s})} \quad (2)$$

Now define $r_{0,0} = m_{0000}(s_0, s)$, $r_{0,2} = m_{0011}(s_0, s)$, $r_{1,0} = m_{0100}(s_0, s) + m_{1000}(s_0, s)$, $r_{1,2} = m_{0111}(s_0, s) + m_{1011}(s_0, s)$, $r_{2,0} = m_{1100}(s_0, s)$, $r_{2,2} = m_{1111}(s_0, s)$. Note that the r 's are dyad counts where the first index represents the sum of the node values and the second index represents the sum of the link values. The above expression can be rewritten as:

$$L(\theta, \beta; d) = P(s | y_s, a_{s_{0U}}) \theta_0^{n_0(s)} (1 - \theta_0)^{n_1(s)} \beta_0^{r_{0,2}} (1 - \beta_0)^{r_{0,0}} \beta_1^{r_{1,2}} (1 - \beta_1)^{r_{1,0}} \beta_2^{r_{2,2}} (1 - \beta_2)^{r_{2,0}} \left[\theta_0 (1 - \beta_0)^{n_0(s_0)} (1 - \beta_1)^{n_1(s_0)} + (1 - \theta_0) (1 - \beta_1)^{n_0(s_0)} (1 - \beta_2)^{n_1(s_0)} \right]^{n(\bar{s})} \quad (3)$$

In the remainder of this paper, we focus on the full symmetric model to illustrate the proposed Bayesian methodology for simplicity of presentation. The same method can be applied to the general model with the likelihood function given in (1).

3.2 Choice of Prior Distributions

Since there are no specific constraints on $\theta_0, \beta_0, \beta_1, \beta_2$, we may assume independent priors on $\theta_0, \beta_0, \beta_1, \beta_2$, all of which take values in the interval $[0, 1]$. It is quite common to put a beta prior on a parameter that takes values in $[0, 1]$ because most smooth unimodal distributions on $[0, 1]$ can be well approximated by some beta distributions and the class of beta distributions is reasonably rich to model the uncertainty about the parameter. Also, the expression in (3) is in general quite complex but beta priors can yield a tractable posterior distribution (to be shown later). Using beta priors, we obtain an analytic formula for the Bayes estimates and the marginal posterior distribution.

In this paper we consider independent beta priors for the parameters:

$$\pi(\theta_0, \beta_0, \beta_1, \beta_2) \propto \theta_0^{a-1} (1 - \theta_0)^{b-1} \beta_0^{c-1} (1 - \beta_0)^{d-1} \beta_1^{e-1} (1 - \beta_1)^{f-1} \beta_2^{g-1} (1 - \beta_2)^{h-1} \quad (4)$$

When determining the constants a and b it is often useful to equate the mean $E[\theta_0] = a/(a+b)$ of $\text{Beta}(a, b)$ to a value which represents your belief about the location of θ_0 and the variance $\text{Var}[\theta_0] = ab/(a+b)^2(a+b+1)$ of $\text{Beta}(a, b)$ to a value which represents the uncertainty put on the specified θ_0 value. Similarly, the values of c, d, e, f, g and

h can be determined. For example, if one is interested in the prevalence of injection drug use in a certain community, one may take an initial sample and trace links by asking the injection drug user in the sample to name the people with whom they share injection equipment. If the value $y_u = 1$ represents injection drug use, then θ_0 is the percentage of non-users in that community. Quite often an estimate for the central location and the spread of θ_0 may be provided.

In the case of complete ignorance, we will consider three commonly used noninformative priors and provide a comparison of the resulting Bayes estimates in our illustrative example in section 4. (For a fuller discussion of the noninformative priors, see Berger 1985, pages 89-90). The first one is the uniform prior, which corresponds to Beta(1,1). The second one, Beta(0, 0), suggested by Haldane (1931), has an improper density. It is equivalent to a prior uniform in the log-odds $\log\{\theta_0/(1-\theta_0)\}$. A possible compromise between Beta(1,1) and Beta(0,0) is Beta(1/2, 1/2), which has a proper density. This prior implies a uniform prior for $\sin^{-1}\sqrt{\theta_0}$.

3.3 Posterior Distribution and Bayes estimates

In our problem, the posterior distribution $\pi(\theta_0, \beta_0, \beta_1, \beta_2 | d)$ corresponding to the beta priors is given by:

$$\begin{aligned} \pi(\theta_0, \beta_0, \beta_1, \beta_2 | d) &\propto \theta_0^{n_0(s)+a-1} (1-\theta_0)^{n_1(s)+b-1} \\ &\quad \beta_0^{r_{0,2}+c-1} (1-\beta_0)^{r_{0,0}+d-1} \\ &\quad \beta_1^{r_{1,2}+e-1} (1-\beta_1)^{r_{1,0}+f-1} \\ &\quad \beta_2^{r_{2,2}+g-1} (1-\beta_2)^{r_{2,0}+h-1} \\ &\quad \left[\theta_0 (1-\beta_0)^{n_0(s_0)} (1-\beta_1)^{n_1(s_0)} \right. \\ &\quad \quad \left. + (1-\theta_0)(1-\beta_1)^{n_0(s_0)} \right]^{n(\bar{s})} \\ &\quad \quad (1-\beta_2)^{n_1(s_0)} \Big]^{n(\bar{s})} \end{aligned} \quad (5)$$

To find the posterior mean (Bayes estimate) of θ_0 , let

$$\begin{aligned} q(\theta_0, \beta_0, \beta_1, \beta_2) &= \theta_0^{n_0(s)+a-1} (1-\theta_0)^{n_1(s)+b-1} \\ &\quad \beta_0^{r_{0,2}+c-1} (1-\beta_0)^{r_{0,0}+d-1} \\ &\quad \beta_1^{r_{1,2}+e-1} (1-\beta_1)^{r_{1,0}+f-1} \\ &\quad \beta_2^{r_{2,2}+g-1} (1-\beta_2)^{r_{2,0}+h-1} \\ &\quad \left[\theta_0 (1-\beta_0)^{n_0(s_0)} (1-\beta_1)^{n_1(s_0)} \right. \\ &\quad \quad \left. + (1-\theta_0)(1-\beta_1)^{n_0(s_0)} \right. \\ &\quad \quad \left. (1-\beta_2)^{n_1(s_0)} \right]^{n(\bar{s})} \end{aligned}$$

Since $\int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = B(\alpha, \beta)$ is the beta function, we have the following two results:

$$\begin{aligned} M_1 &= \int_0^1 \int_0^1 \int_0^1 \int_0^1 q(\theta_0, \beta_0, \beta_1, \beta_2) d\theta_0 d\beta_0 d\beta_1 d\beta_2 \\ &= \sum_{i=0}^{n(\bar{s})} \binom{n(\bar{s})}{i} B(n_0(s)+a+i, n(\bar{s})+n_1(s)+b-i) \\ &\quad B(r_{0,2}+c, i n_0(s_0)+r_{0,0}+d) B(r_{1,2} \\ &\quad +e, i n_1(s_0)+(n(\bar{s})-i) n_0(s_0)+r_{1,0}+f) \\ &\quad B(r_{2,2}+g, (n(\bar{s})-i) n_1(s_0)+r_{2,0}+h). \end{aligned}$$

$$\begin{aligned} M_2 &= \int_0^1 \int_0^1 \int_0^1 \int_0^1 \theta_0 q(\theta_0, \beta_0, \beta_1, \beta_2) d\theta_0 d\beta_0 d\beta_1 d\beta_2 \\ &= \sum_{i=0}^{n(\bar{s})} \binom{n(\bar{s})}{i} B(n_0(s)+a+1+i, n(\bar{s})+n_1(s)+b-i) \\ &\quad B(r_{0,2}+c, i n_0(s_0)+r_{0,0}+d) B(r_{1,2} \\ &\quad +e, i n_1(s_0)+(n(\bar{s})-i) n_0(s_0)+r_{1,0}+f) \\ &\quad B(r_{2,2}+g, (n(\bar{s})-i) n_1(s_0)+r_{2,0}+h). \end{aligned}$$

The Bayes estimate for θ_0 can thus be evaluated by the quotient of the righthand side of the above two equations since:

$$\begin{aligned} E(\theta_0 | d) &= \frac{\int_0^1 \int_0^1 \int_0^1 \int_0^1 \theta_0 q(\theta_0, \beta_0, \beta_1, \beta_2) d\theta_0 d\beta_0 d\beta_1 d\beta_2}{\int_0^1 \int_0^1 \int_0^1 \int_0^1 q(\theta_0, \beta_0, \beta_1, \beta_2) d\theta_0 d\beta_0 d\beta_1 d\beta_2} \\ &= \frac{M_2}{M_1}. \end{aligned}$$

Similarly, the Bayes estimates for $\beta_0, \beta_1, \beta_2$ can be computed.

3.4 Prediction of Realized Graph Quantities

Consider the problem of estimating or predicting, from the sample data, the realized value of some graph quantity $Z = Z(\mathbf{Y}, \mathbf{A})$, an observable but unobserved finite-population quantity. Denoting the unknown parameters collectively by ψ , the relevant posterior predictive density is

$$\begin{aligned} f(z | d) &= \int f(z | d, \psi) \pi(\psi | d) d\psi \\ &\propto \int f(d, z | \psi) \pi(\psi) d\psi \end{aligned} \quad (6)$$

where the constant of proportionality is, as usual, $f(d)$.

For example, suppose the objective is to predict the proportion of nodes in the population that have value $y = 1$. Let $n_1(s)$ denote the number of nodes for which $y = 1$ in the sample, and let $n_1(\bar{s})$ denote the number of nodes with value 1 among the nodes not in the sample. Note that

$n_1(s)$ is observed and $n_1(\bar{s})$ is an unobserved quantity to be estimated or predicted. The realized proportion of value-1 nodes in the population is denoted $Z = (n_1(s) + n_1(\bar{s}))/N$, where N is the total number of nodes in the population.

For a one-wave snowball design with an ignorable initial sample from which all links are traced and with the nondirected stochastic block model, the joint predictive likelihood is

$$f(d, n_1(\bar{s}) | \theta_0, \beta_0, \beta_1, \beta_2) =$$

$$\begin{aligned} & p(s | y_s, a_{s_0}) \binom{n(\bar{s})}{n_1(\bar{s})} \\ & \theta_0^{n_0(s) + n_0(\bar{s})} (1 - \theta_0)^{n_1(s) + n_1(\bar{s})} \\ & \beta_0^{r_{02}(s_0, s)} (1 - \beta_0)^{r_{00}(s_0, s) + n_0(s_0) n_0(\bar{s})} \\ & \beta_1^{r_{12}(s_0, s)} (1 - \beta_1)^{r_{10}(s_0, s) + n_0(s_0) n_1(\bar{s}) + n_1(s_0) n_0(\bar{s})} \\ & \beta_2^{r_{22}(s_0, s)} (1 - \beta_2)^{r_{20}(s_0, s) + n_1(s_0) n_1(\bar{s})}. \end{aligned} \quad (7)$$

With joint likelihood (7) and independent beta priors and carrying out the integration, the posterior predictive density for the finite-population proportion Z becomes

$$\begin{aligned} f(n_1(\bar{s}) | d) & \propto \binom{n(\bar{s})}{n_1(\bar{s})} B[n_0(s) + n_0(\bar{s}) + a, n_1(s) + n_1(\bar{s}) + b] \\ & B[r_{02} + c, r_{00} + n_0(s_0) n_0(\bar{s}) + d] \\ & B[r_{12} + e, r_{10} + n_0(s_0) n_1(\bar{s}) + n_1(s_0) n_0(\bar{s}) + f] \\ & B[r_{22} + g, r_{20} + n_1(s_0) n_1(\bar{s}) + h]. \end{aligned}$$

The Bayes predictor of $n_1(\bar{s})$ is

$$E[n_1(\bar{s}) | d] = \sum_{n_1(\bar{s})=0}^{n(\bar{s})} n_1(\bar{s}) f(n_1(\bar{s}) | d).$$

$$\text{Since } i \binom{n}{i} = n \binom{n-1}{i-1},$$

$$\begin{aligned} E[n_1(\bar{s}) | d] & \propto n(\bar{s}) \sum_{i=1}^{n(\bar{s})} \binom{n(\bar{s})-1}{i-1} B[n_0(s) + n(\bar{s}) - i \\ & + a, n_1(s) + i + b] \\ & B[r_{02} + c, r_{00} + n_0(s_0) (n(\bar{s}) - i) + d] \\ & B[r_{12} + e, r_{10} + n_0(s_0) i + n_1(s_0) (n(\bar{s}) - i) + f] \\ & B[r_{22} + g, r_{20} + n_1(s_0) i + h] \\ & = M_3. \end{aligned}$$

in which M_3 is defined to be the right hand side. Thus, since $M_1 = f(d)$ defined earlier is the proportionality constant, $E[n_1(\bar{s}) | d] = M_3/M_1$.

Therefore, the Bayes predictor \hat{Z} of the realized proportion Z of positive nodes in the population is

$$\begin{aligned} \hat{Z} & = E(Z | d) = E[(n_1(s) + n_1(\bar{s}))/N | d] \\ & = \frac{n_1(s) + (M_3/M_1)}{N}. \end{aligned} \quad (8)$$

4. AN ILLUSTRATIVE EXAMPLE

Here, we consider an example which concerns estimating the percentages of injection drug users and nonusers among a certain target population. Let θ_0 represent the proportion of non injection drug users in the target population. Then $1 - \theta_0$ is the proportion of injection drug users. Suppose that there are 200 people in that population. In the first wave sample, 22 people are sampled randomly without replacement and 5 of those sampled are injection drug users whereas 17 are not. The injection drug users are asked to name their injection partners. Note that links are only possible between users and tracing these links can only add users to the sample. The initial users give 12 referrals, of which 10 are distinct users not in the initial sample. The statistics are:

$$n_1(s_0) = 5, n_0(s_0) = 17, n_1(s) = 15, n_0(s) = 17,$$

$$n(\bar{s}) = 168, r_{22} = 12, r_{20} = 93.$$

In terms of the notation of section 3, $\beta_0 = \lambda_{0011}$ is the probability of a mutual link between two non injection drug users. $\beta_1 = \lambda_{1011} = \lambda_{0111}$ is the probability of a mutual link between injection drug user and non injection drug user (it is natural that the two different orders of node values have the same probability). $\beta_2 = \lambda_{1111}$ is the probability of a mutual link between two injection drug users. Since non injection drug users will by definition not have injection partners, $\beta_0 = \beta_1 = 0$ for this example.

The Bayes estimates for θ_0 and β_2 corresponding to different noninformative priors are given in table 1.

Note that the three noninformative priors are very different from each other. For example, the improper non-informative prior corresponding to $a = b = g = h = 0$ place a lot of its weight on both 0 and 1. This would arise in practice when people in a certain neighbourhood are either all injection drug users or are all non injection drug users, but we just do not know which one. On the other hand, the prior corresponding to $a = b = g = h = 1$ place a flat weight to values between 0 and 1. Even though the three priors are very different, the posterior distributions corresponding to these three non-informative priors nearly coincide with each other. Figure 1 shows the posterior distribution of θ_0 and β_2 corresponding to the three non-informative priors. One can conclude that the Bayes estimates here are not sensitive to the specification of the three priors.

Table 1
Bayes estimates for noninformative priors corresponding to the specified values of a, b, g, h
(The values in the brackets are the 95% HPD regions)

Bayes estimate	$a = b = g = h = 0$	$a = b = g = h = .5$	$a = b = g = h = 1$
$\hat{\theta}_0$.7273 (.5706, .8713)	.7285 (.5747, .8670)	.7295 (.5786, .8686)
$\hat{\beta}_2$.0420 (.0153, .0738)	.0439 (.0164, .0766)	.0458 (.0175, .0791)

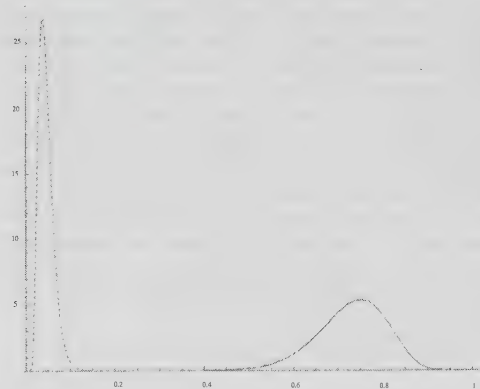


Figure 1. Marginal Posterior distributions: solid line for θ_0 and dashed line for β_2 . (The posterior distributions corresponding to the three non-informative priors are given here and they nearly coincide)

For comparison purposes, it is of interest to note that the maximum likelihood estimates obtained using the likelihood function given in (3) are calculated to be: $\hat{\theta}_0 = .7604$, $\hat{\beta}_2 = .0501$, not far from the Bayes estimates. However, it is not easy to compute confidence intervals for the maximum likelihood estimate whereas one can obtain the posterior intervals for the Bayes estimates without any additional difficulty. For example, a $(1 - \alpha)$ highest posterior density (HPD) region can be obtained for the specified α value for each parameter $\theta_0, \beta_0, \beta_1, \beta_2$, where HPD is the region of values that contains $(1 - \alpha)$ of the posterior probability for that parameter with the characteristic that the density within the region is never lower than that outside. It is worthwhile to note that the posterior intervals can be directly regarded as having the stated probability of containing the unknown quantity in contrast to the repeated sampling property of frequentist confidence interval. See Gelman, Carlin, Stern and Rubin (1995, pages 104-106) for a discussion on the frequency property of some Bayesian procedures.

From Table 1, we can see that even though the width of the HPD interval of β_2 is large compared to the magnitude of its Bayes estimate, it gives us a rough order-of-magnitude estimate of β_2 and provides useful information to the subject matter specialists.

5. AN EMPIRICAL EXAMPLE AND DISCUSSION

To examine the properties of estimators and predictors under repeated sampling, socially-networked data from the Colorado Springs study on the heterosexual transmission of HIV/AIDS was used as an empirical population from which to repeatedly sample. The Colorado Springs study, which is described in Potterat, Woodhouse, Rothenberg, Muth, Darrow, Muth and Reynolds (1993); Rothenberg, Woodhouse, Potterat, Muth, Darrow and Klovdahl (1995), and Darrow, Potterat, Rothenberg, Woodhouse, Muth and Klovdahl (1999), involved a very thorough investigation of a population of people thought to be at high risk for infection with the human immunodeficiency virus. In the study, data were obtained not only on the risk-related behaviors of individuals, but also on their social relationships with other individuals. Risk-related behaviors included various sexual and drug-use behaviors, and the social links examined included sexual and drug-use relationships. Over the course of the study, data were obtained on several thousand people.

For our empirical population we have used the 595 individuals in the study for which the data on both individual risk-related behaviors and relationships to other people in the study are complete. For the node variable of interest we chose a high-risk sexual behavior (commercial sex work) and sexual relationship for the link variable of interest. Figure 2 shows a graphical representation of the empirical population, in which the nodes or circles represent people in the study and the lines represent sexual relationships between pairs of individuals. Presence of the high-risk sexual behavior ($y = 1$) is indicated by a dark colored circle, while presence of a sexual relationship between two individuals is indicated by a line between the two circles. The positioning of the nodes in the graph is arbitrary, but has been arranged to separate connected components. The largest connected component contains 219 of the 595 people in the population. The next largest connected component contains 12 people, followed by several components of 4, 3 and 2 people. There are 267 people without sexual relationships to others among the 595 in the empirical population. The extremely uneven distribution of connected component sizes exemplified by this population presents one of the challenges to sampling design and inference in such populations.

population

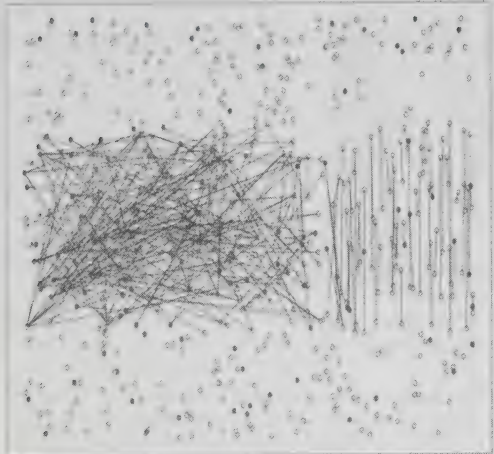


Figure 2. Colorado Springs study on the heterosexual transmission of HIV/AIDS (Potterat *et al.* 1991; Rothenberg *et al.* 1993; Darrow *et al.* 1999): The 595 people in the empirical population. Dark circles represent individuals with high-risk sexual behavior (sex work). Links between circles indicate sexual relationships.

Figure 3. shows a one-wave snowball sample from this population. First, a simple random sample of 40 nodes (circled in the figure) is selected. All links from these initial nodes are traced to add the additional nodes to the sample.

Repeated sampling of the empirical population was carried out using the one-wave snowball design with initial simple random sample of 40 individuals. The addition of a wave of new nodes brought the total sample size to 85, on average. For each sample, various estimators of the proportion of high-risk individuals ($y = 1$) in the population were computed, and this procedure was repeated 1,000 times. The undirected stochastic block graph model was used for the maximum likelihood and Bayes estimators of θ and the Bayes predictor of the finite-population proportion z . A uniform prior was used for the Bayes procedures. Table 2 and Figure 4 summarize the properties under the repeated sampling of the different estimators. The actual proportion of nodes having value ($y = 1$) in the empirical population is 0.2235. The sample proportion overestimates relative to the actual proportion because the linktracing has a tendency to enrich the sample with high-risk nodes. Each of the model-based estimators has relatively little bias with the link-tracing design.

one-wave snowball sample

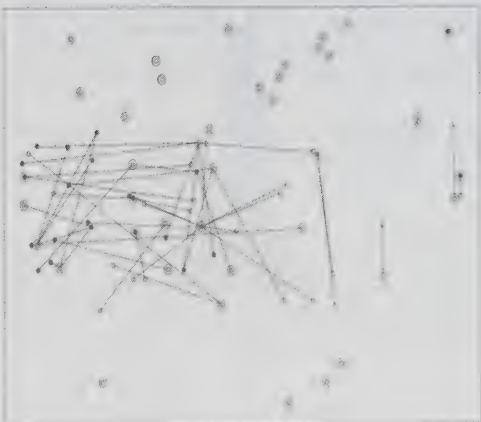


Figure 3. A one-wave snowball sample selected from the Colorado Springs empirical population. From an initial random sample of 40 individuals (circled), links are traced to add one wave of new individuals to the sample.

Table 2

Means and mean square errors of estimators of the population mean of the node values, for the Colorado Springs empirical population. The actual mean of node values in the population is 0.2235294. The design is a one-wave snowball sample with an initial random sample of 40 nodes. The average final sample size was 82.65. The number of simulation runs is 1,000.

Type of estimator:	sample proportion	m.l.e.	Bayes estimator	Bayes predictor
mean:	0.3147	0.2155	0.251	0.2142
m.s.e.:	0.011391	0.003279	0.003261	0.003275

In this paper, we employ a Bayesian approach to the estimation problem with link-tracing design and show that, corresponding to the independent beta priors, the posterior distribution can be evaluated analytically. If a more general prior is desired then one can use the Markov Chain Monte Carlo (MCMC) method to evaluate the posterior for that general prior. References for using MCMC techniques in Bayesian computations include Gilks, Richardson and Spiegelhalter (1996) and Gelman, Carlin, Stern and Rubin (1995). The approach used in Gelfand and Smith (1990) can be adapted for the implementation of the MCMC simulations here.

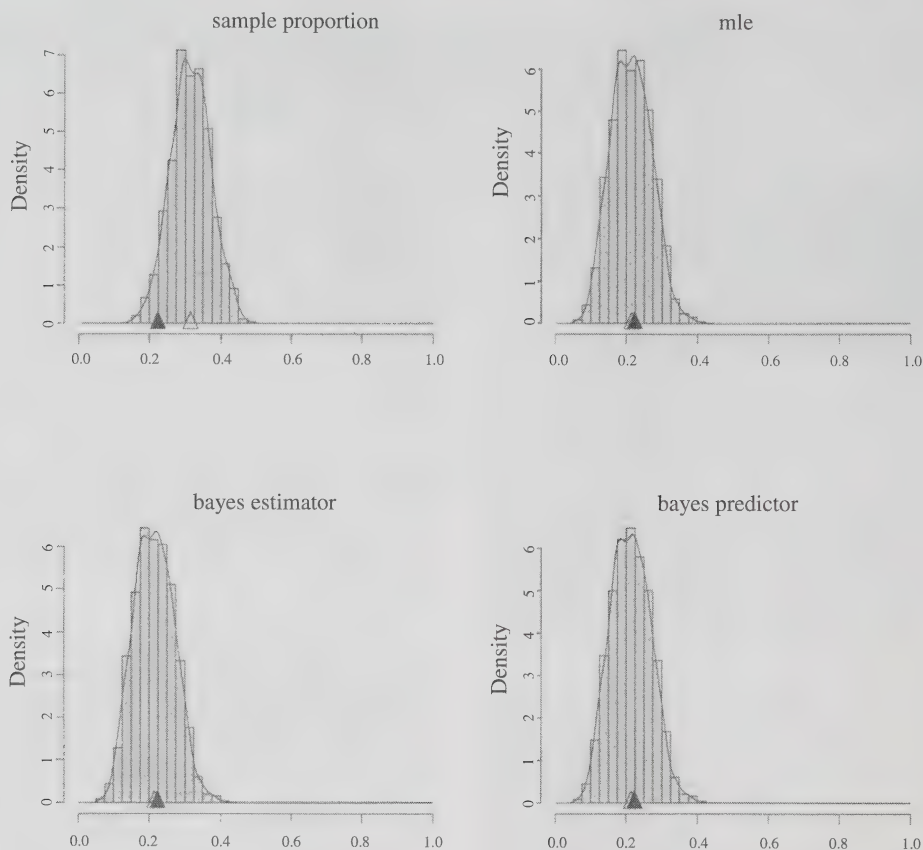


Figure 4. Distributions of estimators of the proportion of individuals in the high-risk category in the Colorado Springs empirical population, with the one-wave snowball design using an initial sample of 40. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. The number of simulations was 1000.

ACKNOWLEDGEMENTS

Support for this work was provided by funding from the National Center for Health Statistics, the National Science Foundation (DMS-9626102), and the National Institutes of Health (R01-DA09872). The authors would like to thank John Potterat and Steve Muth for advice and use of the data from the Colorado Springs study. We would also like to thank the Associated Editor and the referees for their insightful comments and suggestions.

REFERENCES

- BERGER, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, (2nd ed.). Berlin: Springer-Verlag.
- DARROW, W.W., POTTERAT, J.J., ROTHENBERG, R.B., WOODHOUSE, D.E., MUTH, S.Q. and KLOVDAHL, A.S. (1999). Using knowledge of social networks to prevent human immunodeficiency virus infections: The Colorado Springs Study. *Sociological Focus*. 32, 143-158.
- ERICKSON, B. (1978). Some problems of inference from chain data. In *Sociological Methodology*, 1979, K.F. Schuessler (Ed.) San Francisco: Jossey-Bass. 276-302.
- FRANK, O. (1971). *Statistical Inference in Graphs*. Stockholm: Försvarets forskningsanstalt.
- FRANK, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference*. 1, 235-246.
- FRANK, O. (1977b). A note on Bernoulli sampling in graphs and Horvitz-Thompson estimation. *Scandinavian Journal of Statistics*. 4, 178-180.

- FRANK, O. (1977c). Estimation of graph totals. *Scandinavian Journal of Statistics*. 4, 81-89.
- FRANK, O. (1978). Estimating the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*. 5, 177-188.
- FRANK, O. (1979). Estimation of population totals by use of snowball samples. In *Perspectives on Social Network Research*. (P.W. Holland, and S. Leinhardt, Eds.). New York: Academic Press. 319-348.
- FRANK, O. (1980). Sampling and inference in a population graph. *International Statistical Review*. 48, 33-41.
- FRANK, O. (1997). Composition and structure of social networks. *Mathematiques, Informatique et Sciences humaines*. 35, 11-23.
- FRANK, O., and SNIJDERS, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*. 10, 53-67.
- GELFAND, A.E., and SMITH, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*. 85, 398-409.
- GELMAN, A., CARLIN, J., STERN, H. and RUBIN, D (1995). *Bayesian Data Analysis*. London: Chapman & Hall.
- GILK, W.R, RICHARDSON, S. and SPIEGELHALTER (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.
- GOODMAN, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*. 20, 572-579.
- HALDANE, J.B.S. (1931). A note on inverse probability. *Proc Cambridge Philos. Soc*. 28, 55-61.
- NEAIGUS, A., FRIDEMAN, S.R., GOLDSTEIN, M.F., ILDEFONSO, G., CURTIS, R. and JOSE, B. (1995). Using dyadic data for a network analysis of HIV infection and risk behaviors among injection drug users. In *Social Networks, Drug Abuse, and HIV Transmission*. (R.H. Needle, S.G. Genser and R.T. II Trotter, Eds.) NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse. 20-37.
- NEAIGUS, A., FRIEDMAN, S.R., JOSE, B., GOLDSTEIN, M.F., CURTIS, R., ILDEFONSO, G. and DES JARLAIS, D.C. (1996). High-risk personal networks and syringe sharing as risk factors for HIV infection among new drug injectors. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*. 11, 499-509.
- POTTERAT, J.J., WOODHOUSE, D.E., ROTHENBERG, R.B., MUTH, S.Q., DARROW, W.W., MUTH, J.B. and REYNOLDS, J.U. (1993). AIDS in Colorado Springs: Is there an epidemic? *AIDS*. 7, 1517-1521.
- ROTHENBERG, R.B., WOODHOUSE, D.E., POTTERAT, J.J., MUTH, S.Q., DARROW, W.W. and KLOVDAHL, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. In *Social Networks, Drug Abuse, and HIV Transmission*. (R.H. Needle, S.G. Genser and R.T. II Trotter, Eds.) NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse. 3-19.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*. 63, 581-592.
- SNIJDERS, T.A.B. (1992). Estimation on the basis of snowball samples: how to weight. *Bulletin de Methodologie Sociologique*. 36, 59-70.
- SNIJDERS, T.A.B., and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*. 14, 75-100.
- THOMPSON, S.K., and COLLINS, L.M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence*. 68, S57-S67.
- THOMPSON, S.K., and FRANK, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*. 26, 87-98.
- THOMPSON, S.K., and SEBER, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees during 2002. An asterisk indicates that the person served more than once.

- M.Z. Anis, *Indian Statistical Institute*
 J. Bethel, *Westat, Inc.*
 J.-F. Beaumont, *Statistics Canada*
 D.R. Bellhouse, *University of Western Ontario*
 M. Bellow, *NASS*
 Y. Berger, *University of Southampton*
 D. Binder, *Statistics Canada*
 E. Blair, *University of Houston*
 J. Breidt, *Iowa State University*
 J.M. Brick, *Westat, Inc.*
 R. Chambers, *University of Southampton*
 J. Chen, *University of Waterloo*
 M.J. Cho, *Bureau of Labor Statistics*
 J. Choi, *National Center for Health Statistics*
 J. Church, *Worsey House*
 C. Clark, *U.S. Bureau of the Census*
 P. Clarke, *Office for National Statistics*
 R. Clark, *Australian Bureau of Statistics*
 M.P. Cohen, *U.S. Bureau of Transportation Statistics*
 F. Conrad, *University of Michigan*
 M.P. Couper, *University of Michigan*
 F.A. Cowell, *London School of Economics and Political Science*
 J. Dalen, *Eurostat*
 J. de Haan, *Statistics Netherlands*
 P. Dick, *Statistics Canada*
 A. Dorfman, *U.S. Bureau of Labour Statistics*
 P. Duchesne, *Université de Montréal*
 M.R. Elliott, *University of Pennsylvania*
 J. Eltinge, *U.S. Bureau of Labor Statistics*
 W.A. Fuller, *Iowa State University*
 J. Gambino, *Statistics Canada*
 M. Ghosh, *University of Florida*
 A. Gower, *Statistics Canada*
 B. Graubard, *National Cancer Institute*
 S. Hawala, *U.S. Census Bureau*
 D. Haziza, *Statistics Canada*
 D. Hedeker, *University of Illinois*
 D. Hedlin, *University of Southampton*
 D.F. Heitjan, *University of Pennsylvania*
 M.A. Hidioglou, *Statistics Canada*
 S. Hinkins, *National Opinion Research Centre*
 T. Holt, *University of Southampton*
 B. Hulliger, *Swiss Federal Statistical Office*
 D. Judkins, *Westat, Inc.*
 G. Kalton, *Westat Inc.*
 A. Kennickell, *Federal Reserve System*
 J.-K. Kim, *Hankuk University of Foreign Studies*
 * P. Kott, *National Agricultural Statistics Service*
 V. Kuusela, *Statistics Finland*
 P.A. Lachenbruch, *U.S. Food and Drug Administration*
 P. Lahiri, *JPSM, University of Maryland*
 N. Laniel, *Statistics Canada*
 * P. Lavallée, *Statistics Canada*
 * H. Lee, *Westat, Inc.*
 R. Lehtonen, *University of Jyväskylä*
 J. Lent, *U.S. Bureau of Transportation Statistics*
 J. Lepkowski, *University of Michigan*
 R.J.A. Little, *University of Michigan*
 S. Linacre, *Australian Bureau of Statistics*
 S. Lohr, *Arizona State University*
 T. Maiti, *Iowa State*
 H. Mantel, *Statistics Canada*
 S. Matthews, *Statistics Canada*
 X.-L. Meng, *Harvard University*
 S.M. Miller, *U.S. Bureau of Labour Statistics*
 S.R. Mohen, *ISI*
 J.M. Montaquila, *Westat, Inc.*
 G. Nathan, *The Hebrew University of Jerusalem*
 D. Norris, *Statistics Canada*
 * J. Opsomer, *Iowa State University*
 D. Pfeiffermann, *The Hebrew University of Jerusalem*
 N.G.N. Prasad, *University of Alberta*
 * T.E. Raghunathan, *University of Michigan*
 J.N.K. Rao, *Carleton University*
 P.S.R.S. Rao, *University Rochester*
 T.J. Rao, *Indian Statistical Institute*
 J. Reiter, *Duke University*
 L.-P. Rivest, *Université Laval*
 P. Saavedra, *ORC Macro*
 * S. Sae-Ung, *U.S. Census Bureau*
 C.-E. Särndal, *University of Montreal*
 J. Schafer, *Pennsylvania State University*
 N. Schenker, *National Center for Health Statistics*
 F.J. Scheuren, *National Opinion Research Center*
 M.D. Sinclair, *Mathematica Policy research*
 R. Sitter, *Simon Fraser University*
 * C. Skinner, *University of Southampton*
 K.P. Srinath, *ABT Associates*
 E. Stasny, *Ohio State University*
 J.-L. Tambay, *Statistics Canada*
 Y. Thibaudeau, *U.S. Bureau of the Census*
 Y. Tillé, *Université de Neuchâtel*
 R. Valliant, *Westat, Inc.*
 J. van der Brakel, *Statistics Netherlands*
 V. Vehovar, *University of Ljubljana*
 J. Waksberg, *Westat, Inc.*
 W.E. Winkler, *U.S. Bureau of the Census*

K.M. Wolter, *Iowa State University*
C. Wu, *University of Waterloo*
W. Yung, *Statistics Canada*

A. Zaslavsky, *Harvard University*
H. Zheng, *Harvard Medical School*
K. Zieschang, *International Monetary Fund*

Acknowledgements are also due to those who assisted during the production of the 2003 issues: H. Laplante, F. Pilon-Renaud and R. Guido (Dissemination Division) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge C. Cousineau, C. Ethier, and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 19, No. 2, 2003

Weighting Methods Graham Kalton and Ismael Flores-Cervantes	81
Penalized Spline Model-Based Estimation of the Finite Populations Total from Probability-Proportional-to-Size Samples Hui Zheng and Roderick J.A. Little	99
Optimal Calibration Estimators Under Two-Phase Sampling Changbao Wu and Ying Luan	119
A Method for Estimating Design-based Sampling Variances for Surveys with Weighting, Poststratification, and Raking Hao Lu and Andrew Gelman	133
Prevention and Treatment of Item Nonresponse Edith D. de Leeuw, Joop Hox, and Mark Huisman	153
Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics Dan Hedlin	177
Book and Software Reviews	201
In Other Journals	211

Contents Volume 19, No. 3, 2003

Monthly Disaggregation of a Quarterly Time Series and Forecasts of Its Unobservable Monthly Values Victor M. Guerrero	215
A Post-stratified Raking-ratio Estimator Linking National and State Survey Data for Estimating Drug Use Trent D. Buskirk and Jane L. Meza	237
Simultaneous Estimation of the Mean of a Binary Variable from a Large Number of Small Areas Li-Chun Zhang	253
A Practical Use for Instrumental-Variable Calibration Phillip S. Kott	265
Exploring the Meaning of Consent: Participation in Research and Beliefs about Risks and Benefits Eleanor Singer	273
Quality Issues at Statistics Norway Hans Viggo Saeboe, Jan Byfuglien, and Randi Johannessen	287
book and Software Reviews	305

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

Volume 31, No. 2, June/juin 2003, 115-238

Nicole MALFAIT & James O. RAMSAY The historical functional linear model	115
Sujit K. SAHU, Dipak K. DEY & Márcia D. BRANCO A new class of multivariate skew distributions with application to Bayesian regression models	129
Chunming M. ZHANG Adaptive test of regression functions via multiscale generalized likelihoods ratios	151
Gerda CLAESKENS, Bing-Yi JING, Liang PENG & Wang ZHOU Empirical likelihood confidence regions for comparison distributions and ROC curves	173
Yann GUÉDON & Christiane COCOZZA-THIVENT Nonparametric estimation of renewal processes from count data	191
Inna PEREVOZSKAYA, William F. ROSENBERGER & Linda M. HAINES Optimal design for the proportional odds model	225
Forthcoming Papers/Articles à paraître	237
Volume 31 (2003): Subscription rates/Frais d'abonnement	238

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in Word or WordPerfect. A paper copy may be required for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points ci-dessous. Les articles acceptés doivent être soumis sous forme de fichiers de traitement de texte, préféralement Word ou WordPerfect. Une version papier pourrait être requise pour les formules et les graphiques.

1. Présentation

- 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$ etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 0; 1).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Volume 31, No. 2, June/juin 2003, 115-238

Nicole MALPAIT & James O. RAMSAY The historical functional linear model	115
Sujit K. SAHU, Dipak K. DEY & Márcia D. BRANCO A new class of multivariate skew distributions with application to Bayesian regression models	129
Chunming M. ZHANG Adaptive test of regression functions via multiscale generalized likelihoods ratios	151
Gerda CLAESKENS, Bing-Yi JING, Liang PENG & Wang ZHOU Empirical likelihood confidence regions for comparison distributions and ROC curves	173
Yann GUÉDON & Christiane COCOZZA-THIVENT Nonparametric estimation of renewal processes from count data	191
Inna PEREVOZSKAYA, William F. ROSENBERGER & Linda M. HAINES Optimal design for the proportional odds model	225
Forthcoming Papers/Articles à paraître	237
Volume 31 (2003): Subscription rates/Frais d'abonnement	238

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 19, No. 2, 2003

81	Weighting Methods Graham Kalton and Ismael Flores-Cervantes
99	Penalized Spline Model-Based Estimation of the Finite Populations Total from Probability-Proportional-to-Size Samples Hui Zheng and Roderick J.A. Little
119	Optimal Calibration Estimators Under Two-Phase Sampling Changbao Wu and Ying Luan
133	A Method for Estimating Design-based Sampling Variances for Surveys with Weighting, Poststratification, and Raking Hao Lu and Andrew Gelman
153	Prevention and Treatment of Item Nonresponse Edith D. de Leeuw, Joop Hox, and Mark Huisman
177	Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics Dan Hedlin
201	Book and Software Reviews
211	In Other Journals

Contents Volume 19, No. 3, 2003

215	Monthly Disaggregation of a Quarterly Time Series and Forecasts of Its Unobservable Monthly Values Victor M. Guerrero
237	A Post-stratified Raking-ratio Estimator Linking National and State Survey Data for Estimating Drug Use Trent D. Buskirk and Jane L. Meza
253	Simultaneous Estimation of the Mean of a Binary Variable from Large Number of Small Areas Li-Chun Zhang
265	A Practical Use for Instrumental-Variable Calibration Phillip S. Kott
273	Exploring the Meaning of Consent: Participation in Research and Beliefs about Risks and Benefits Eleanor Singer
287	Quality Issues at Statistics Norway Hans Viggo Saeboe, Jan Byfuglien, and Randi Johannessen
305	book and Software Reviews

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

K.M. Wolter, *Iowa State University*
 C. Wu, *University of Waterloo*
 W. Yung, *Statistique Canada*

A. Zaslavsky, *Harvard University*
 H. Zheng, *Harvard Medical School*
 K. Zieschang, *International Monetary Fund*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 2003: H. Laplante, F. Pilon-Renaud et R. Guido (Division de la diffusion) et L. Perrault (Division des langues officielles et traduction). Finalement on désire exprimer notre reconnaissance à C. Cousineau, C. Ethier et D. Lemire de la Division des méthodes des enquêtes auprès des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article durant l'année 2002. Un astérisque indique que la personne a participé plus d'une fois.

- V. Kuusela, *Statistics Finland*
 P.A. Lachenbruch, *U.S. Food and Drug Administration*
 P. Lahiri, *JPSM, University of Maryland*
 N. Laniel, *Statistique Canada*
 P. Lavallée, *Statistique Canada* *
- H. Lee, *Westat, Inc.* *
- R. Lehtonen, *University of Jyväskylä*
 J. Lent, *U.S. Bureau of Transportation Statistics*
 J. Lepkowski, *University of Michigan*
 R.J.A. Little, *University of Michigan*
 S. Linacre, *Australian Bureau of Statistics*
 S. Lohr, *Arizona State University*
 T. Maiti, *Iowa State*
 H. Mantel, *Statistics Canada*
 S. Matthews, *Statistique Canada*
 X.-L. Meng, *Harvard University*
 S.M. Miller, *U.S. Bureau of Labour Statistics*
 S.R. Mohe, *ISI*
 J.M. Montaquilla, *Westat, Inc.*
 N.G.N. Prasad, *University of Alberta*
 G. Nathan, *The Hebrew University of Jerusalem*
 D. Norris, *Statistique Canada* *
- J. Opsomer, *Iowa State University* *
- D. Pfeffermann, *The Hebrew University of Jerusalem* *
- T.E. Raghunathan, *University of Michigan* *
- J.N.K. Rao, *Carleton University*
 P.S.R.S. Rao, *University Rochester*
 T.J. Rao, *Indian Statistical Institute*
 J. Reiter, *Duke University*
 L.-P. Rivest, *Université Laval*
 P. Saavedra, *ORC Macro* *
- S. Sae-Ung, *U.S. Census Bureau* *
- C.-E. Sämäl, *University of Montreal*
 J. Schafer, *Pennsylvania State University*
 N. Schenker, *National Opinion Research Center*
 F.J. Scheuren, *National Opinion Research Center*
 M.D. Sinclair, *Mathematica Policy research* *
- C. Skinner, *University of Southampton* *
- K.P. Srinath, *ABT Associates*
 E. Stasny, *Ohio State University*
 J.-L. Tambay, *Statistique Canada*
 Y. Thibaut, *U.S. Bureau of the Census*
 Y. Tillé, *Université de Neuchâtel*
 R. Valliant, *Westat, Inc.*
 J. van der Brakel, *Statistics Netherlands*
 V. Vechov, *University of Ljubljana*
 J. Waksberg, *Westat, Inc.*
 W.E. Winkler, *U.S. Bureau of the Census*
- M.Z. Anis, *Indian Statistical Institute*
 J. Bethel, *Westat, Inc.*
 J.-F. Beaumont, *Statistique Canada*
 D.R. Bellhouse, *University of Western Ontario*
 M. Bellow, *NAASS*
 Y. Berger, *University of Southampton*
 D. Binder, *Statistique Canada*
 E. Blair, *University of Houston*
 J. Breidt, *Iowa State University*
 J.M. Brick, *Westat, Inc.*
 R. Chambers, *University of Southampton*
 J. Chen, *University of Waterloo*
 M.J. Cho, *Bureau of Labour Statistics*
 J. Choi, *National Center for Health Statistics*
 J. Church, *Worsley House*
 C. Clark, *U.S. Bureau of the Census*
 P. Clarke, *Office for National Statistics*
 R. Clark, *Australian Bureau of Statistics*
 M.P. Cohen, *U.S. Bureau of Transportation Statistics*
 F. Conrad, *University of Michigan*
 M.P. Couper, *University of Michigan*
 F.A. Cowell, *London School of Economics and Political Science*
- J. Dalen, *Eurostat*
 J. de Haan, *Statistics Netherlands*
 P. Dick, *Statistique Canada*
 A. Dorfman, *U.S. Bureau of Labour Statistics*
 P. Duchesne, *Université de Montréal*
 M.R. Elliott, *University of Pennsylvania*
 J. Ethinge, *U.S. Bureau of Labour Statistics*
 W.A. Fuller, *Iowa State University*
 J. Gambino, *Statistique Canada*
 M. Ghosh, *University of Florida*
 A. Gower, *Statistique Canada*
 B. Graubard, *National Cancer Institute*
 S. Hawala, *U.S. Census Bureau*
 D. Haziza, *Statistique Canada*
 D. Hedeker, *University of Illinois*
 D. Hedlin, *University of Pennsylvania*
 D.F. Heijman, *University of Pennsylvania*
 M.A. Hidiroglou, *Statistique Canada*
 S. Hinkins, *National Opinion Research Centre*
 T. Holt, *University of Southampton*
 B. Hultinger, *Swiss Federal Statistical Office*
 D. Judkins, *Westat, Inc.*
 G. Kalton, *Westat Inc.*
 A. Kennickell, *Federal Reserve System*
 J.-K. Kim, *Hankuk University of Foreign Studies* *
- F. Kott, *National Agricultural Statistics Service*

- FRANK, O. (1980). Sampling and inference in a population graph. *Revue Internationale de Statistique*, 48, 33-41.
- FRANK, O. (1997). Composition and structure of social networks. *Mathématiques, Informatique et Sciences humaines*, 35, 11-23.
- FRANK, O., et SNIJDERS, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.
- GELFAND, A.E., et SMITH, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.
- GELMAN, A., CARLIN, J., STERN, H. et RUBIN, D. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.
- GILK, W.R., RICHARDSON, S. et SPIEGELHALTER (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.
- GOODMAN, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 20, 572-579.
- HALDANE, J.B.S. (1931). A note on inverse probability. *Proc Cambridge Philos. Soc.* 28, 55-61.
- NEAIGUS, A., FRIDEMAN, S.R., GOLDSTEIN, M.F., ILDEFONSO, G., CURTIS, R. et JOSE, B. (1995). Using dyadic data for a network analysis of HIV infection and risk behaviors among injection drug users. Dans *Social Networks, Drug Abuse, and HIV Transmission*, (R.H. Needle, S.G. Gensser et R.T. II Trotter, Eds.) NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 20-37.
- NEAIGUS, A., FRIDEMAN, S.R., JOSE, B., GOLDSTEIN, M.F., CURTIS, R., ILDEFONSO, G. et DES JARLAIS, D.C. (1996). High-risk personal networks and syringe sharing as risk factors for HIV infection among new drug injectors. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 11, 499-509.
- FRANK, O. (1980). Sampling and inference in a population graph. *Revue Internationale de Statistique*, 48, 33-41.
- FRANK, O. (1997). Composition and structure of social networks. *Mathématiques, Informatique et Sciences humaines*, 35, 11-23.
- FRANK, O., et SNIJDERS, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.
- GELFAND, A.E., et SMITH, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.
- GELMAN, A., CARLIN, J., STERN, H. et RUBIN, D. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.
- GILK, W.R., RICHARDSON, S. et SPIEGELHALTER (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.
- GOODMAN, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 20, 572-579.
- HALDANE, J.B.S. (1931). A note on inverse probability. *Proc Cambridge Philos. Soc.* 28, 55-61.
- NEAIGUS, A., FRIDEMAN, S.R., GOLDSTEIN, M.F., ILDEFONSO, G., CURTIS, R. et JOSE, B. (1995). Using dyadic data for a network analysis of HIV infection and risk behaviors among injection drug users. Dans *Social Networks, Drug Abuse, and HIV Transmission*, (R.H. Needle, S.G. Gensser et R.T. II Trotter, Eds.) NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 20-37.
- NEAIGUS, A., FRIDEMAN, S.R., JOSE, B., GOLDSTEIN, M.F., CURTIS, R., ILDEFONSO, G. et DES JARLAIS, D.C. (1996). High-risk personal networks and syringe sharing as risk factors for HIV infection among new drug injectors. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 11, 499-509.
- THOMPSON, S.K., et SEBER, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.
- THOMPSON, S.K., et FRANK, O. (2000). Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépistage de liens. *Techniques d'enquête*, 26, 99-112.
- THOMPSON, S.K., et COLLINS, L.M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence*, 68, S57-S67.
- THOMPSON, S.K., et NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14, 75-100.
- SNIJDERS, T.A.B., et NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14, 75-100.
- SNIJDERS, T.A.B. (1992). Estimation on the basis of snowball sample: how to weigh. *Bulletin de Méthodologie Sociologique*, 36, 59-70.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- RUBIN, D.B. (1992). Estimation on the basis of snowball sample: how to weigh. *Bulletin de Méthodologie Sociologique*, 36, 59-70.
- ROTHENBERG, R.B., WOODHOUSE, D.E., POTTERAT, J.J., MUTH, S.Q., DARROW, W.W., et KLOVDADHL, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. Dans *Social Networks, Drug Abuse, and HIV Transmission*, (R.H. Needle, S.G. Gensser et R.T. II Trotter, Eds.) NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 3-19.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- SNIJDERS, T.A.B. (1992). Estimation on the basis of snowball sample: how to weigh. *Bulletin de Méthodologie Sociologique*, 36, 59-70.
- SNIJDERS, T.A.B., et NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14, 75-100.
- THOMPSON, S.K., et COLLINS, L.M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence*, 68, S57-S67.
- THOMPSON, S.K., et FRANK, O. (2000). Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépistage de liens. *Techniques d'enquête*, 26, 99-112.
- THOMPSON, S.K., et SEBER, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.

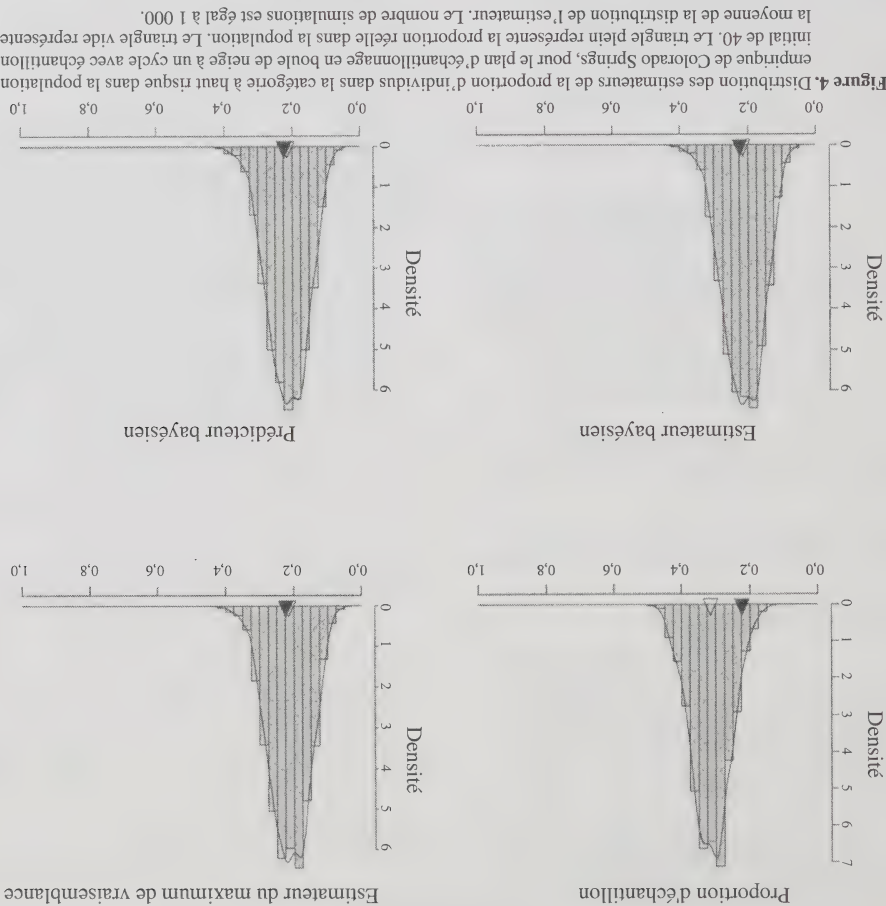


Figure 4. Distribution des estimateurs de la proportion d'individus dans la catégorie à haut risque dans la population empirique de Colorado Springs, pour le plan d'échantillonnage en boule de neige à un cycle avec échantillon initial de 40. Le triangle plein représente la proportion réelle dans la population. Le triangle vide représente la moyenne de la distribution de l'estimateur. Le nombre de simulations est égal à 1 000.

REMERCIEMENTS

Les travaux ont été financés par le National Center for Health Statistics, la National Science Foundation (DMS-9626102) et les National Institutes of Health (R01-DA09872). Nous remercions John Poterat et Steve Muth de leurs conseils et de nous avoir permis d'utiliser les données de l'étude de Colorado Springs. Nous remercions aussi le rédacteur adjoint et les examinateurs de leurs commentaires et suggestions constructifs.

BIBLIOGRAPHIE

- BERGER, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, (2^e éd.). Berlin: Springer-Verlag.
- DARROW, W.W., POTTERAT, J.J., ROTHENBERG, R.B., WOODHOUSE, D.E., MUTH, S.Q. et KLOVDAH, A.S. (1999). Using knowledge of social networks to prevent human immunodeficiency virus infections: The Colorado Springs Study. *Sociological Focus*, 32, 143-158.
- FRANK, O. (1979). Estimation of population totals by use of snowball samples. In *Perspectives on Social Network Research*, (P.W. Holland, et S. Leinhardt, Eds.). New York: Academic Press, 319-348.
- FRANK, O. (1978). Estimating the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*, 5, 177-188.
- FRANK, O. (1977c). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4, 178-180.
- FRANK, O. (1977b). A note on Bernoulli sampling in graphs and Horvitz-Thompson estimation. *Scandinavian Journal of Statistics*, 4, 178-180.
- FRANK, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-246.
- FRANK, O. (1971). *Statistical Inference in Graphs*. Stockholm: Försvarets forskningsanstalt.
- FRANK, O. (1971). *Statistical Inference in Graphs*. Stockholm: Francis & Taylor.
- FRANK, O. (1978). Some problems of inference from chain data. Dans *Sociological Methodology*, 1979, (K.F. Schuessler Ed.). San Francisco: Jossey-Bass, 276-302.

Nous avons procédé à l'échantillonnage répété de la population empirique selon un plan d'échantillonnage simple boule de neige à un cycle avec échantillon aléatoire simple initial de 40 individus. L'ajout d'un cycle de nouveaux nœuds a donné une taille totale d'échantillon de 85, en moyenne. Pour chaque échantillon, nous avons calculé divers estimateurs de la proportion d'individus à haut risque ($y = 1$) et répété cette procédure 1 000 fois. Nous avons utilisé le modèle de graphe en bloc non orienté si- chastique pour l'estimateur du maximum de vraisemblance et l'estimateur bayésien de θ , ainsi que le prédicteur bayésien de la proportion en population π . Nous avons utilisé une loi a priori uniforme pour les méthodes bayésiennes. Le tableau 2 et la figure 4 résument les propriétés sous échantillonnage répété de divers esti- mateurs. La proportion réelle de nœuds ayant la valeur ($y = 1$) dans la population empirique est 0,2235. La proportion d'échantillon est une surestimation comparati- vement à la proportion réelle, parce que le dépiégeage des liens a tendance à enrichir l'échantillon en nœuds à haut risque. Dans le cas du plan d'échantillonnage par dépiégeage de liens, le biais de chaque estimateur fondé sur un modèle est assez faible.

Tableau 2
Moyennes et erreurs quadratiques moyennes des estimateurs de la moyenne de population des valeurs nodales, pour la population empirique de Colorado Springs. La moyenne réelle des valeurs nodales dans la population est 0,2235294. Le plan d'échantillonnage est un plan en boule de neige à un cycle avec échantillon aléatoire initial de 40 nœuds. La taille moyenne de l'échantillon final est 82,65. Le nombre de simulations exécutées est 1 000.

Type	Proportion	Estimateur du maximum de vrai- semblance	Estimateur bayésien	Prédicteur bayésien
moyenne	0,3147	0,2155	0,251	0,2142
eqm	0,011391	0,003279	0,003261	0,003275

Dans le présent article, nous proposons une approche bayésienne du problème d'estimation en cas de plan d'échantillonnage par dépiégeage de liens et montrons que pour chaque distribution indépendante bêta a priori, on peut évaluer analytiquement la loi a posteriori cor- respondante. Si l'on souhaite une loi a priori plus générale, on peut utiliser une méthode de Monte Carlo à chaîne de Markov (MCMC) pour évaluer la loi a posteriori. Les ré- férences concernant l'utilisation des techniques MCMC en calcul bayésien incluent Gills, Richardson et Spiegelhalter (1996) et Gelman, Carlin, Stern et Rubin (1995). L'approche de Gelfand et Smith (1990) peut être adaptée pour des simulations MCMC.

sélectionné un échantillon aléatoire simple de 40 nœuds (encercles dans la figure). Tous les liens à partir de ces nœuds initiaux sont tracés pour ajouter les nœuds supplémentaires à l'échantillon.

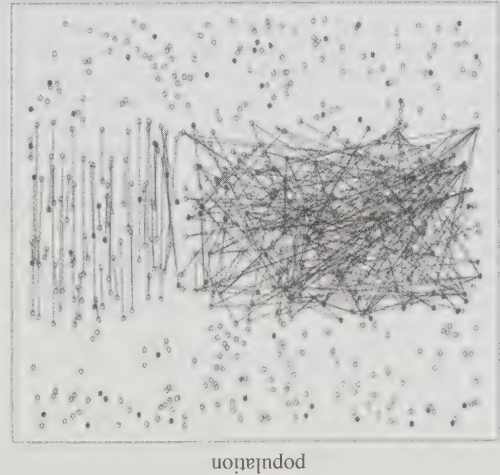


Figure 2. Étude de Colorado Springs sur la transmission hétérosexuelle du VIH/SIDA (Pottarat et coll. 1991; Rothenberg et coll. 1993; Darrow et coll. 1999) : Les cercles foncés représentent les individus présentant un comportement sexuel à haut risque (prostitution). Les liens entre les cercles indiquent les relations sexuelles.

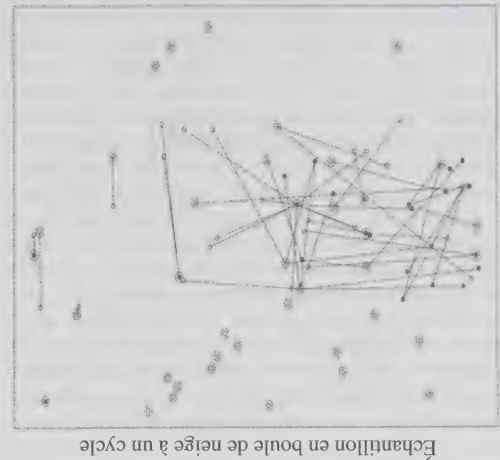


Figure 3. Échantillon en boule de neige à un cycle tiré de la population empirique de Colorado Springs. Pour un échantillon aléatoire initial de 40 individus (encercles), on a tracé les liens pour ajouter un cycle de nouveaux individus à l'échantillon.

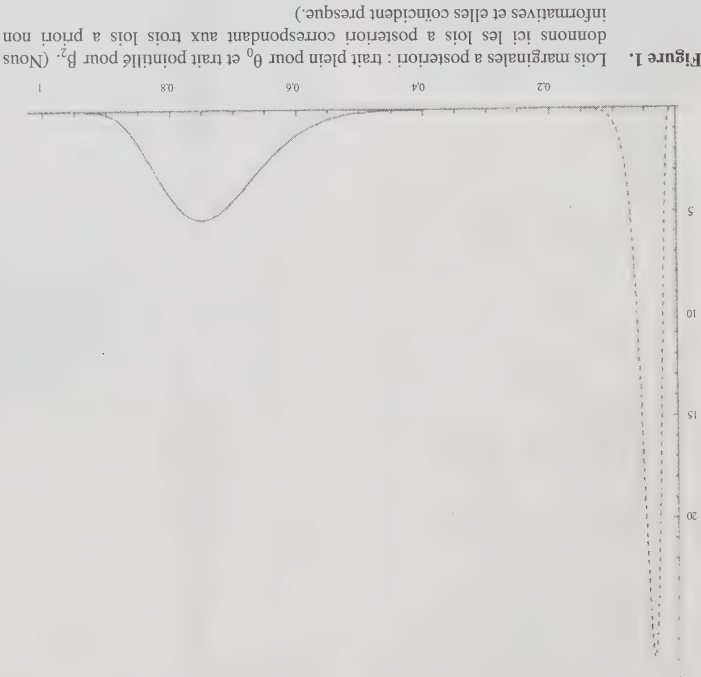


Figure 1. Lois marginales a posteriori : trait plein θ_0 et trait pointillé pour β_2 . (Nous donnons ici les lois a posteriori correspondant aux trois lois a priori non informatives et elles coïncident presque.)

5. EXEMPLE EMPIRIQUE ET DISCUSSION

Pour examiner les propriétés des estimateurs et des prédicteurs sous échantillonnage répété, nous avons utilisé les données obtenues par dépistage des liens sociaux lors de l'étude de Colorado Springs sur la transmission hétérosexuelle du VIH/SIDA comme population empirique à partir de laquelle a été effectué l'échantillonnage répété. Darrow, Muth et Reynolds (1993); Rothenberg, Muth, Darrow, Potterat, Muth, Darrow et Klov Dahl (1995), et Klov Dahl (1999), l'étude de Colorado Springs représente une évaluation très minutieuse d'une population de personnes considérées comme courant un risque élevé d'infection par le virus de l'immunodéficience humaine. Des données ont été recueillies non seulement sur les comportements à risque des individus, mais aussi sur leurs relations sociales avec d'autres individus. Les comportements à risque incluent divers comportements sexuels et de consommation de drogues, et les liens sociaux étudiés incluent les relations sexuelles et de consommation de drogues. Durant l'étude, des données ont été recueillies auprès de plusieurs milliers de personnes.

Nous avons utilisé comme population empirique les 595 individus ayant participé à l'étude pour lesquels les données sur les comportements à risque individuels et les relations avec d'autres personnes incluses dans l'étude sont complètes. Comme variable nodale d'intérêt, nous avons choisi un comportement sexuel à haut risque (prostitution) et comme variable de lien étudiée, nous avons choisi une relation sexuelle. La figure 2 donne une représentation graphique de la population empirique, où les nœuds ou cercles représentent les personnes participant à l'étude et les lignes représentent les relations sexuelles entre pères d'individus. La présence d'un comportement sexuel à haut risque ($y = 1$) est indiquée par un cercle foncé, tandis que la présence d'une relation sexuelle entre deux individus est indiquée par une ligne entre les deux cercles. Le positionnement des nœuds dans le graphique est arbitraire et a été arrangé de façon à séparer les composantes connectées. La composante connectée la plus importante contient 219 des 595 personnes formant la population. La composante connectée la plus importante qui suit compte 12 personnes. Vient ensuite plusieurs composantes de 4, 3 et 2 personnes. Dans la population empirique de 595 personnes, 267 n'ont pas de relations sexuelles avec d'autres. La répartition extrêmement inégale des tailles des composantes connectées illustre par cette population l'illustration d'un plan d'échantillonnage et l'inférence dans le cas de telles populations.

La figure 3 montre un échantillon en boule de neige à un cycle de cette population. Pour commencer, nous avons

Notons que les trois lois a priori non informatives sont très différentes les unes des autres. Par exemple, la loi a priori non informative incorrecte correspondant à $a = b = g = h = 0$ accorde beaucoup de poids aux valeurs 0 et 1. Cela se produit, en pratique, si les personnes vivant dans un quartier particulier sont soit toutes des utilisateurs soit toutes des non-utilisateurs de drogues injectables, mais que l'on ne sait pas ce qu'elles sont. En revanche, la loi a priori correspondant à $a = b = g = h = 1$ accorde un poids uniforme aux valeurs comprises entre 0 et 1. Bien que les trois lois a priori soient fort différentes, les lois a posteriori correspondant à ces trois lois a priori non informatives coïncident presque. La figure 1 montre la loi a posteriori de θ_0 et β_2 correspondant aux trois lois a priori non informatives. Nous pouvons conclure qu'ici, les estimations bayésiennes ne sont pas sensibles à la spécification des trois lois a priori.

Aux fins de comparaison, il est intéressant de noter que les estimations du maximum de vraisemblance obtenues au moyen de la fonction de vraisemblance donnée en (3) sont : $\theta_0 = 0,7604$, $\beta_2 = 0,0501$, valeurs qui s'écartent peu des estimations bayésiennes. Cependant, il n'est pas facile de calculer les intervalles de confiance pour les estimations du maximum de vraisemblance, alors qu'on peut obtenir les intervalles a posteriori pour les estimations bayésiennes sans difficulté supplémentaire. Par exemple, on peut obtenir une région à densité a posteriori la plus élevée (DPPE) $(1 - \alpha)$ pour la valeur de α spécifiée pour chaque paramètre $\theta_0, \beta_0, \beta_1, \beta_2$, où la région DPPE est la région de valeurs qui contient $(1 - \alpha)$ de la probabilité a posteriori pour le paramètre en question, caractérisée par le fait que la densité n'est jamais plus faible à l'intérieur qu'à l'extérieur de la région. Il mérite d'être souligné que les intervalles a posteriori peuvent être considérés directement comme ayant la probabilité énoncée de contenir la quantité inconnue, contrairement à la propriété d'échantillonnage répété de l'intervalle de confiance fréquentiste. Consulter Gelman, Carlin, Stern et Rubin (1995, pages 104 à 106) pour une discussion de la propriété fréquentiste de certaines procédures bayésiennes.

L'examen du tableau 1 montre que, même si l'intervalle DPPE de β_2 est grand comparativement à la grandeur de l'estimation bayésienne correspondante, il donne une idée grossière de l'ordre de grandeur de β_2 et fournit des renseignements utiles aux spécialistes du domaine.

Tableau 1
Estimations bayésiennes pour les lois a priori non informatives correspondant aux valeurs spécifiées de a, b, g, h
(Les valeurs entre parenthèses sont les régions DPPE à 95 %)

Estimation bayésienne			
$a = b = g = h = 0$	$a = b = g = h = 0,5$	$a = b = g = h = 1$	
0,7273 (0,5706, 0,86713)	0,7285 (0,5747, 0,8670)	0,7295 (0,5786, 0,8686)	θ_0
0,0420 (0,0153, 0,0738)	0,0439 (0,0164, 0,0766)	0,0458 (0,0175, 0,0791)	β_2

Par conséquent, le prédicteur bayésien Z de la proportion réalisée Z de nœuds positifs dans la population est

$$Z = E(Z|d) = E[n_1(s) + n_1(\bar{s})]/N|d| \tag{8}$$
$$= \frac{n_1(s) + (M_3/M_1)N}{n_1(s) + (M_3/M_1)N}$$

4. UN EXEMPLE

Ici, nous considérons un exemple d'estimation des drogues injectables et de non-utilisateurs de drogues injectables parmi une population cible particulière. Soit θ_0 la proportion de non-utilisateurs de drogues injectables dans la population cible. Alors, $1 - \theta_0$ est la proportion d'utilisateurs de drogues injectables. Supposons que la population compte 200 personnes. Durant le premier cycle d'échantillonnage, 22 personnes sont sélectionnées aléatoirement sans remise et cinq de ces personnes échantillonnées sont des utilisateurs de drogues injectables, tandis que 17 ne le sont pas. On demande aux utilisateurs de drogues injectables de nommer les personnes avec lesquelles ils prennent ces drogues. Notons que les liens ne sont possibles qu'entre utilisateurs et que le suivi de ces liens permet uniquement d'ajouter des utilisateurs à l'échantillon. Les utilisateurs initiaux nomment 12 personnes, dont 10 sont des utilisateurs distincts ne figurant pas dans l'échantillon initial. Les chiffres sont :

$$n_1(s_0) = 5, n_0(s_0) = 17, n_1(s) = 15, n_0(s) = 17, n(\bar{s}) = 168, r_{22} = 12, r_{20} = 93.$$

Selon la notation utilisée à la section 3, $\beta_0 = \lambda^{001}$ est la probabilité d'un lien mutuel entre deux non-utilisateurs de drogues injectables, $\beta_1 = \lambda^{101}$ est la probabilité d'un lien mutuel entre un utilisateur et un non-utilisateur de drogues injectables (il est naturel que les deux ordres distincts de valeurs nodales ait la même probabilité). $\beta_2 = \lambda^{111}$ est la probabilité d'un lien mutuel entre deux utilisateurs de drogues injectables. Puisque, par définition, les non-utilisateurs n'ont pas de partenaires avec qui ils prennent des drogues injectables, $\beta_0 = \beta_1 = 0$ dans l'exemple considéré.

Les estimations bayésiennes de θ_0 et β_2 correspondant à diverses lois a priori non informatives sont données au tableau 1.

dans l'échantillon et soit $n_1(s)$ le nombre de nœuds dont la valeur est 1 parmi les nœuds ne figurant pas dans l'échantillon. Notons que $n_1(s)$ est observé et que $n_1(s)$ est une quantité inobservée qu'il faut estimer ou prédire. La proportion réalisée de nœuds de valeur 1 dans la population est représentée par $Z = (n_1(s) + n_1(s)')/N$, où N est le nombre total de nœuds dans la population.

$$\text{conjoint est } f(d, n_1(\bar{s}) | \theta^0, \beta^0, \beta^1, \beta^2) =$$

On peut donc obtenir l'estimation bayésienne de θ_0 en calculant le quotient du deuxième membre des deux équations susmentionnées, puisque :

Nous pouvons calculer les estimations bayésiennes de β_0 , β_1 et β_2 de la même façon.

Considérons le problème de l'estimation ou de la prévision, à partir des données d'échantillon, de la valeur

$$\mathbb{E} p(p | \mathbb{A}) \mathbb{E}(\mathbb{A} | p | z) f \int = (p | z) f$$

où la constante de proportionnalité est, comme d'habitude,

ou M_3 est défini comme étant le deuxième membre. Donc, puisque $M_1 = f(d)$ défini antérieurement est la constante de proportionnalité, $E[n_1(\bar{s}) | d] = M_3/M_1$.

ces drogues contenus dans l'échantillon de nommer les personnes avec qui ils ont partagé du matériel d'injection. Si la valeur $y^n_n = 1$ représente l'utilisation de drogues injectables, alors θ_0 est le pourcentage de non-utilisateurs dans cette collectivité. Assez souvent, une estimation de l'emplacement central et de l'étendue de θ_0 peut être fournie.

Dans le cas où l'on peut omettre complètement de tenir compte du plan d'échantillonnage, nous considérons trois lois a priori non informatives, voir Berger 1985, pages 89 et 90). La première est la loi a priori uniforme, qui correspond à bêta (1,1). La deuxième, bêta (0,0), proposée par Haldane (1931), a une densité incorrecte. Elle est équivalente à une loi a priori uniforme en log-odds $\log(\theta_0/(1-\theta_0))$. Un compromis possible entre bêta (1,1) et bêta (0,0) est bêta (1/2, 1/2), dont la densité est correcte. Cette loi a priori sous-entend une loi a priori uniforme pour $\sin^{-1}\sqrt{\theta_0}$.

3.3 Loi a posteriori et estimations bayésiennes

Dans notre problème, la loi a posteriori $\pi(\theta_0, \beta_0, \beta_1, \beta_2 | d)$ correspondant aux lois a priori bêta est donnée par :

$$\pi(\theta_0, \beta_0, \beta_1, \beta_2 | d) \propto \theta_0^{n_0(s)-1} (1-\theta_0)^{n_1(s)+d-1} \times \beta_0^{r_{0,0}-1} (1-\beta_0)^{r_{0,0}+d-1} \times \beta_1^{r_{1,0}-1} (1-\beta_1)^{r_{1,0}+f-1} \times \beta_2^{r_{2,0}-1} (1-\beta_2)^{r_{2,0}+h-1} \times \left[\theta_0^d (1-\theta_0)^{n_0(s)-d} (1-\beta_1)^{n_1(s)} (1-\beta_2)^{n_2(s)} \right] \times \left[(1-\theta_0)(1-\beta_1) \right]^{n(s)}$$

Pour trouver la moyenne a posteriori (estimation bayésienne) de θ_0 , posons que

$b(\theta_0, \beta_0, \beta_1, \beta_2) = \theta_0^{n_0(s)+a-1} (1-\theta_0)^{n_1(s)+d-1} \times \beta_0^{r_{0,0}+c-1} (1-\beta_0)^{r_{0,0}+d-1} \times \beta_1^{r_{1,0}+e-1} (1-\beta_1)^{r_{1,0}+f-1} \times \beta_2^{r_{2,0}+g-1} (1-\beta_2)^{r_{2,0}+h-1} \times \left[\theta_0^d (1-\beta_0)^{n_0(s)-d} (1-\beta_1)^{n_1(s)} (1-\beta_2)^{n_2(s)} \right] \times \left[(1-\theta_0)(1-\beta_1) \right]^{n(s)}$

Définissons maintenant $r_{0,0} = m_{0000}(s_0, s), r_{0,2} = m_{0010}(s_0, s), r_{1,0} = m_{0100}(s_0, s), r_{1,2} = m_{0011}(s_0, s), r_{2,0} = m_{1000}(s_0, s), r_{2,2} = m_{1111}(s_0, s)$. Notons que les r sont des nombres de dyades, où le premier indice représente la somme des valeurs modales et le deuxième, la somme des valeurs de lien. L'expression qui précède peut être réécrite sous la forme :

$$L(\theta, \beta, d) = p(s | y^s, a^{s_0}) \theta_0^{n_0(s)} (1-\theta_0)^{n_1(s)} \beta_0^{r_{0,0}} (1-\beta_0)^{r_{0,0}} \times \beta_1^{r_{1,0}} (1-\beta_1)^{r_{1,0}} \beta_2^{r_{2,0}} (1-\beta_2)^{r_{2,0}} \times \left[\theta_0^d (1-\beta_0)^{n_0(s_0)-d} (1-\beta_1)^{n_1(s_0)} (1-\beta_2)^{n_2(s_0)} \right] \times \left[(1-\theta_0)(1-\beta_1) \right]^{n(s_0)}$$

Dans la suite de l'article, pour la simplicité de l'exposé, nous nous concentrons sur le modèle symétrique complet en vue d'illustrer la méthode bayésienne proposée. La même méthode peut être appliquée au modèle général avec la fonction de vraisemblance donnée par (1).

3.2 Choix des lois a priori

Puisqu'il n'existe aucune contrainte particulière sur $\theta_0, \beta_0, \beta_1, \beta_2$, nous pouvons supposer que les lois a priori de $\theta_0, \beta_0, \beta_1, \beta_2$ sont indépendantes et qu'elles prennent toutes les valeurs comprises dans l'intervalle [0, 1]. Il est assez courant d'attribuer une loi a priori bêta à un paramètre qui prend les valeurs comprises dans [0, 1], parce qu'on peut obtenir une bonne approximation de la plupart des distributions lisses unimodales sur [0, 1] au moyen de certaines lois bêta et que la catégorie des lois bêta est suffisamment riche pour modéliser l'incertitude au sujet du paramètre. En outre, l'expression (3) est, en général, assez complexe, mais les lois a priori bêta peuvent produire une loi a posteriori calculable (à montrer plus tard). Au moyen de lois a priori bêta, nous obtenons une formule analytique pour les estimations bayésiennes et la loi marginale a posteriori.

Dans le présent article, nous considérons des lois a priori bêta pour les paramètres :

$\pi(\theta_0, \beta_0, \beta_1, \beta_2) \propto \theta_0^{a-1} (1-\theta_0)^{b-1} \beta_0^{c-1} (1-\beta_0)^{d-1} \times \beta_1^{e-1} (1-\beta_1)^{f-1} \beta_2^{g-1} (1-\beta_2)^{h-1} \quad (4)$

Quand on détermine les constantes a et b , il est souvent utile d'égaliser la moyenne $E[\theta_0] = a/(a+b)$ de bêta (a, b) à une valeur qui représente la croyance au sujet de l'emplacement de θ_0 et la variance $\text{Var}[\theta_0] = ab/(a+b)^2(a+b+1)$ de bêta (a, b) à une valeur qui représente l'incertitude attachée à la valeur de θ_0 spécifiée. On peut déterminer de la même façon les valeurs telles que c, d, e, f, g et h . Par exemple, pour déterminer la prévalence de la consommation de drogues injectables dans une collectivité particulière, on pourrait tirer un échantillon initial et dépister les liens en demandant aux utilisateurs de

3. INFÉRENCE BAYÉSIENNE À PARTIR DE PLANS D'ÉCHANTILLONNAGE PAR DÉPISTAGE DE LIEN

3.1 Fonction de vraisemblance étant donné les données d'échantillon

Un échantillon s du graphe est un sous-ensemble de nœuds provenant de U et un sous-ensemble de paires de nœuds provenant de U^2 . Les données d'échantillon $d = (s, y_s, a_{s0})$ sont fonction de l'échantillon sélectionné et des valeurs graphiques y et a . Pour tout plan d'échantillonnage dans lequel la sélection de l'échantillon dépend des valeurs graphiques y et a uniquement par la voie dépend des valeurs y_s et a_s incluses dans les données, le plan d'échantillonnage n'influe pas sur la valeur des estimateurs de vraisemblance ou les estimateurs bayésiens (Rubin 1976, Thompson et Frank 2000). Ainsi, nombre de plans d'échantillonnage en boucle de nœuds et d'autres plans d'échantillonnage par dépistage de liens sont négligeables en cas d'inférence fondée sur la vraisemblance, à condition que la méthode de sélection de l'échantillon initial soit négligeable. Dans ce sens, tout plan d'échantillonnage convenable, ou adaptatif soigneusement mis en œuvre serait négligeable. On peut obtenir des échantillons initiaux négligeables de non négligeables lorsque le tirage n'est pas contrôlé et que les probabilités de sélection sont reliées à des valeurs de nœuds et de liens non observées, par exemple dans le cas où des personnes évitant de prendre des risques et ne comptant qu'un petit nombre de relations sont moins facilement repérées par les enquêteurs, donc exercent une influence non mesurée sur les unités qui sont manquantes, donc sur les probabilités de sélection dans l'échantillon.

Considérons le plan d'échantillonnage par dépistage de liens conformément auquel on sélectionne un échantillon initial s_0 puis on suit tous les liens partant de nœuds compris dans s_0 pour ajouter l'ensemble s_1 de nœuds ne figurant pas dans s_0 qui sont adjacents aux nœuds compris dans s_0 . L'échantillon complet est $s = s_0 \cup s_1$. L'ensemble complet d'étiquettes dans la population peut être représenté par l'union de tous ensembles disjoints, $U = s_0 \cup s_1 \cup \bar{s}$, où \bar{s} représente les nœuds non échantillonnés. Ici, nous considérons un plan d'échantillonnage dans lequel la décision de suivre les liens à partir d'un nœud u dépend de la valeur nodale y_u . Par exemple, dans une étude de l'utilisation de drogues injectables, l'échantillon initial pourrait contenir à la fois des utilisateurs et des non-utilisateurs. Si les chercheurs décident de suivre les liens sociaux uniquement à partir des utilisateurs, le plan d'échantillonnage dépend adaptativement des valeurs nodales y , ainsi que des liens. Le plan d'échantillonnage peut alors être représenté par $P(s|y_s, a_{s0})$, puisque la méthode de sélection dépend à la fois des valeurs des nœuds et des liens. Les données

sont $d = (s, y_s, a_{s0})$. Puisque la décision ne dépend des valeurs y et a que par la voie des données observées, les paramètres du plan d'échantillonnage s'éliment de la fonction de vraisemblance, par mise en facteurs, et de la loi de Bayes à posteriori, par division, si bien que l'inférence fondée sur la vraisemblance ou l'inférence bayésienne dépend uniquement du modèle supposé.

Pour le modèle de graphe décrit à la section précédente, il s'ensuit (Thompson et Frank 2000) que la vraisemblance avec les données d'échantillon est :

$$L(\theta, \lambda; d) = P(s|y_s, a_{s0}) \sum_{N=1}^n \left(\prod_{i=1}^N \theta_{\lambda_i} \right) \left(\prod_{i=N+1}^n \lambda_{\lambda_i} \right) \left(\prod_{i=N+1}^n \lambda_{\lambda_i}^{a_{s0,i}} \right)$$

où la somme est calculée sur toutes les valeurs de y^n et a^n qui ne sont pas fixées par les données d'échantillon.

Pour les plans d'échantillonnage par dépistage de liens dans lesquels on suit tous les liens, plutôt qu'un sous-échantillon de ceux-ci, à partir des nœuds de l'échantillon initial, tous les éléments de la sous-matrice a_{s0} sont nuls. Thompson et Frank (2000) ont montré que la fonction de

$$L(\theta, \lambda; Y, A) = P(s|Y_s, A_{s0}) \left(\prod_{i=1}^I \theta_{\lambda_i} \right) \left(\prod_{i=1}^I \lambda_{\lambda_i}^{a_{s0,i}} \right) \left(\prod_{i=1}^I \lambda_{\lambda_i}^{a_{s0,i}} \right)$$

$$\times \left[\sum_{j=1}^f \theta_{\lambda_j} \prod_{i=1}^I \lambda_{\lambda_i}^{a_{s0,i}} \right] \left(\prod_{i=1}^I \lambda_{\lambda_i}^{a_{s0,i}} \right)$$

(1)

où $n(s)$, $n_i(s_0)$ et $n_i(\bar{s})$ représentent les nombres de nœuds de type i dans l'échantillon complet s , l'échantillon initial s_0 et les nœuds non échantillonnés \bar{s} , respectivement, et $m_{ijkl}(s_0, s_1)$, $m_{ijkl}(s_0, \bar{s})$ sont les nombres de paires de nœuds dans $s_0 \times s_0$ et $s_0 \times \bar{s}$.

Pour un modèle symétrique, $\lambda_{ijkl} = 0$ pour $k \neq l$, de sorte que les arcs sont toujours bidirectionnels ou, de façon équivalente, peuvent être considérés comme des arcs non orientées. Les paramètres du modèle symétrique complet sont $\lambda_{ijkk} = \lambda_{ijik}$, $k = 0, 1$, avec $\lambda_{ij00} + \lambda_{ij11} = 1$. Pour simplifier la notation de ce modèle, posons que β_k représente la probabilité d'un lien mutuel entre deux nœuds dont la valeur totale est k , pour $k = 0, 1$ ou 2 . L'équation de la fonction de vraisemblance susmentionnée se simplifie pour donner

$$L(\theta, \beta; d) = P(s|Y_s, a_{s0}) \left(\prod_{i=1}^I \theta_{\lambda_i} \right) \left(\prod_{i=1}^I \beta_{i+1}^{a_{s0,i}} \right) \left(\prod_{i=1}^I \beta_{i+1}^{a_{s0,i}} \right)$$

(2)

À la section 2, nous donnons la notation pour un modèle de graphe complet contenant les liens associés aux valeurs nodales et sa fonction de vraisemblance. À la section 3, nous présentons la fonction de vraisemblance pour l'échantillon obtenu à partir d'un plan d'échantillonnage par dépistage de liens, ainsi qu'une méthode d'inférence bayésienne. À la section 4, nous donnons un exemple. À la section 5, nous concluons l'article par un exemple empirique et une discussion.

2. LE MODÈLE

Au moyen d'une notation comparable à celle utilisée par Frank (1971) et par Thompson et Frank (2000), nous représentons l'ensemble complet d'étiquettes de nœud par $U = \{1, 2, \dots, N\}$ qui forme la population de N unités. Une variable d'intérêt associée à un nœud individuel u sera représentée par X_u , tandis qu'une variable d'intérêt associée à une paire de nœuds u et v sera représentée par A^{uv} . La série de variables nodales d'intérêt est représentée par $\mathbf{X} = (X_1, \dots, X_N)$. Ici, nous considérons la variable d'intérêt A^{uv} comme étant une variable directionnelle dont la valeur est égale à un s'il existe un arc (lien directionnel) de u à v , et nulle autrement, pour deux nœuds distincts u et v . La matrice des indicateurs d'arc, où A^{uv} est l'élément situé dans la u^{e} ligne et la v^{e} colonne, est la matrice de contiguité du graphe, représentée par \mathbf{A} . Par souci de simplicité, nous supposons que les éléments diagonaux A^{uu} sont nuls. La paire ordonnée (u, v) est décrite comme étant une dyade de type $(X_u, X_v; A^{uv}, A^{vu})$. Dans le modèle hypothétique qui suit, les variables nodales X_1, \dots, X_N sont des variables aléatoires de Bernoulli indépendantes et indépendamment distribuées (i.d.) avec probabilité $P(X_u = 1) = \theta_1^u$, pour $i = 0, 1$ et $\theta_0^u + \theta_1^u = 1$. Conditionnellement aux valeurs nodales X_1^N, \dots, X_N^N , les dyades (A^{uv}, A^{vu}) sont indépendantes pour $1 \leq u < v \leq N$, et leur loi conditionnelle est donnée par $P[(X_u^{uv}, A^{uv}), (X_v^{vu}, A^{vu})] = \binom{k}{i} \binom{l}{j} \lambda_{ij}^{kl} X_u^i X_v^j$ pour toutes les combinaisons de $i = 0, 1, j = 0, 1$, et $k = 0, 1, l = 0, 1$. Pour toutes les combinaisons de i et j , les sommes sur k et l sont représentées par $\lambda_{ij}^{kl} = \sum_k \lambda_{ij}^{kl}$ et égales à 1. Afin d'obtenir des probabilités graphiques ne dépendant pas des identités nodales, nous supposons que sont remplies les conditions de symétrie naturelle suivantes : $\lambda_{0010}^{1110} = \lambda_{0101}^{1011} = \lambda_{0110}^{1101} = \lambda_{0100}^{1001}$ et $\lambda_{0010}^{1000} = \lambda_{0100}^{1010}$. Par exemple, la première et la cinquième conditions indiquent que, entre deux nœuds ayant la même valeur, la probabilité qu'il existe un arc dans la même direction est la même. Soit N_i le nombre total de nœuds de valeur i dans le graphe, de sorte que $N_0 + N_1 = N$. Soit, en outre, M_{ijkl}^{ijkl} le nombre total de dyades de type (i, j, k, l) , c'est-à-dire le nombre total de paires ordonnées de nœuds (u, v) tels que $(X_u^i, X_v^j, A^{uv} = k, A^{vu} = l)$. La fonction de vraisemblance pour le graphe complet sous le modèle avec les paramètres (θ, λ) est

$$L(\theta, \lambda; \mathbf{X}, \mathbf{A}) = (\prod_{N_i=0}^1 \theta_N^i)^N (\prod_{i=0}^1 \prod_{j=0}^1 \prod_{k=0}^1 \prod_{l=0}^1 \lambda_{ijkl}^{M_{ijkl}}).$$

partir d'un point d'origine jusqu'à un point final) peuvent être utilisées en pratique. Snijders (1992) a utilisé la même expression « échantillonnage en boule de neige » en vue d'inclure les plans d'échantillonnage où l'on se limite à suivre un sous-échantillon de liens à partir de chaque nœud. Frank et Snijders (1994) considèrent l'estimation fondée sur un modèle et le plan d'échantillonnage de la taille d'une population cachée – c'est-à-dire le nombre de nœuds dans le graphe – au moyen de l'échantillonnage en boule de neige. Une autre méthode de dépistage des liens pour laquelle il existe des estimateurs fondés sur le plan d'échantillonnage est l'échantillonnage en grappes adaptatif (Thompson et Seber 1996), qui a été formulé dans le contexte graphique ainsi que dans le contexte spatial.

Dans le cas d'une méthode fondée sur le plan d'échantillonnage en population fixe, dans le contexte graphique, on considère les caractéristiques des personnes ainsi que de la structure du réseau social de la population comme étant des valeurs fixes, inconnues. Les propriétés telles que l'absence de biais par rapport au plan d'échantillonnage ne sont fonction d'aucune hypothèse au sujet de la population proprement dite, mais elles dépendent de l'exécution du plan d'échantillonnage tel que spécifié. Dans le présent article, nous considérons des estimateurs fondés sur un modèle, puisqu'elles peuvent être appliquées à une grande gamme de méthodes d'échantillonnage. Très souvent, dans les études de populations cachées et d'accès difficile, on ne peut analyser facilement les résultats des méthodes de sélection d'échantillons, y compris le dépistage des liens, en sondage, mais on peut appliquer les résultats des méthodes fondées sur un modèle à ces cas.

Thompson et Frank (2000) utilisent une méthode fondée sur un modèle pour faire des inférences dans le cas d'un plan d'échantillonnage par dépistage de liens. Dans leur article, ils décrivent des estimateurs du maximum de vraisemblance des paramètres et les prédicteurs de quantités réalisées de graphes de population. Ici, nous adoptons une approche bayésienne pour résoudre le problème d'estimation des graphes. Pour les problèmes réels avec plan d'échantillonnage suivant les liens sociaux d'une personne à l'autre, on pourrait disposer d'information a priori sur les caractéristiques que l'on veut estimer. L'utilisation efficace de cette information par une méthode bayésienne devrait produire de meilleurs estimateurs. En outre, si l'information disponible est vague, on peut utiliser des lois a priori non informatives et effectuer une analyse de sensibilité. Il est important de souligner que, dans le cadre bayésien, on peut obtenir des estimations des intervalles en vue d'évaluer l'exactitude des estimations sans trop de difficulté supplémentaire, alors que cette tâche serait ardue si l'on utilisait la méthode du maximum de vraisemblance. Nous traitons l'inférence concernant les caractéristiques des nœuds ainsi que celles des arcs, comme la prévalence de la maladie dans une collectivité particulière et le taux de transmission de cette maladie d'un sujet à l'autre.

Estimation avec plans d'échantillonnage par dépistage de liens – Une approche bayésienne

MOSUK CHOW et STEVEN K. THOMPSON¹

RÉSUMÉ

L'échantillonnage par dépistage de liens consiste à suivre les liens sociaux d'un répondant à l'autre pour obtenir d'échantillonnage est souvent le seul moyen pratique d'obtenir un échantillon suffisamment grand pour que l'étude donne de bons résultats. Dans le présent article, nous proposons une approche bayésienne du problème d'estimation. Lors des études fondées sur un plan d'échantillonnage par dépistage de liens, on dispose parfois de renseignements à priori sur les caractéristiques des individus. Si l'information disponible est vague, on peut utiliser des lois à priori non informatives pour produire de meilleurs estimateurs. Si l'information disponible est vague, on peut utiliser des lois à priori non informatives et procéder à une analyse de sensibilité. Dans notre exemple, nous constatons que les estimateurs ne sont pas sensibles aux lois à priori spécifiées. Il est important de souligner que, dans le cadre de travail bayésien, l'estimation d'intervalles pour évaluer l'exactitude des estimateurs peut se faire sans difficulté. Par contre, ces estimations sont difficiles à calculer par la méthode classique. En général, une analyse bayésienne donne, pour les paramètres inconnus, une loi (la loi a posteriori) à partir de laquelle il est possible de répondre à un grand nombre de questions simultanément.

MOTS CLÉS : Plans d'échantillonnage par dépistage de liens, échantillonnage en boule de neige, échantillonnage adaptatif, échantillonnage par graphe, échantillonnage de réseau, loi a priori béli.

1. INTRODUCTION

Les données sur les réseaux sociaux comprennent des mesures des relations entre les personnes ou d'autres entités sociales, ainsi que des mesures sur les entités proprement dites. La collecte de données sur des réseaux entiers demandant beaucoup de temps et d'énergie, surtout s'ils sont grands, il est important de pouvoir estimer les propriétés des réseaux d'après des échantillons. Dans les plans d'échantillonnage par dépistage de liens, on suit les liens sociaux d'un répondant à l'autre pour obtenir l'échantillon. Pour les populations humaines cachées et d'accès difficile, ce genre de plan d'échantillonnage est souvent le seul moyen pratique d'obtenir un échantillon suffisamment grand pour que l'étude donne de bons résultats. Par exemple, pour étudier la relation entre l'utilisation de drogues injectables et la propagation de l'infection par le VIH, on peut suivre les liens sociaux mentionnés par les répondants initiaux et ajouter à l'échantillon les individus ainsi dépistés (voir, par exemple, Neaigus, Friedman, Goldstein, Curtis et Jose, 1995; Neaigus, Friedman, Jose, Goldstein, Curtis, Ildelfonso et Des Jarlais 1996; Thompson et Collins 2002). De même, pour étudier des sans-abri, on peut demander aux répondants des renseignements sur d'autres sans-abri qui sont alors échantillonnés. On modélise souvent les populations présentant une structure sociale au moyen de graphes où les nœuds représentent les personnes et les arcs, les liens sociaux, les relations ou les transactions. Dans le contexte graphique, les nœuds et celles associées aux nœuds de nœuds. Le graphe de la population proprement dit peut être considéré comme une structure fixe ou comme une réalisation d'un modèle de graphe stochastique. Des échantillonnages sont sélectionnés pour recueillir des renseignements sur ce graphe. Habituellement, la méthode d'échantillonnage tient compte des arcs ou liens unissant une entité à une autre. Les publications, de nature tant appliquée que théorique, sur l'échantillonnage de réseau sont abondantes. Frank (1977a, 1977b, 1977c, 1978, 1979, 1980, 1997) a obtenu de nombreux résultats importants dans le domaine de l'échantillonnage des réseaux sociaux. Son ouvrage classique (Frank 1971) présente des solutions de base pour l'estimation de quantités graphiques à partir de données d'échantillon. Snijders et Nowicki (1997) proposent diverses méthodes statistiques, y compris une méthode bayésienne, pour l'estimation et la prédiction au moyen de modèles en blocs stochastiques pour les graphes pour lesquels les valeurs nodales ne sont pas observées. L'échantillonnage en boule de neige (Goodman 1961) est une forme de plan d'échantillonnage par dépistage de liens où l'on demande aux individus faisant partie de l'échantillon initial d'identifier des connaissances auxquelles on demande d'identifier, à leur tour, des connaissances, et ainsi de suite pour un nombre fixe d'étapes ou de cycles. Erickson (1978) et Frank (1979) examinent les plans d'échantillonnage en boule de neige dans le but de comprendre comment d'autres « méthodes en chaîne » (méthodes conçues pour dépister les liens dans un réseau à

¹ Mosuk Chow et Steven K. Thompson, Department of Statistics, The Pennsylvania State University, 326 Thomas Building, University Park, PA 16802, États-Unis.

6. SOMMAIRE

Nous établissons des estimateurs pour ce qui, dans le

contexte de toute méthode de sélection à grande entropie, est une bonne approximation de la variance par rapport au plan de sondage de l'estimateur HT du total.

Ces estimateurs ressemblent, mais ne sont pas identiques, à d'autres estimateurs de la variance proposés pour des méthodes de sélection particulièrement à grande entropie par Hájek (1964), Rosen (1997) et Deville (1999). Ces esti-

mateurs offrent tous l'avantage important, comparativement à l'estimateur de la variance type SYG, d'avoir une formule ne comprenant pas les probabilités d'inclusion de deuxième ordre π_{ij} .

Des études empiriques montrent que ces estimateurs ont

tous d'assez bonnes propriétés, à la fois pour le cas spécial important $n = 2$ et lorsque n prend une valeur plus élevée. L'estimateur donne par (16) avec c_j défini par (18), qui

possède certaines propriétés théoriques quasi optimales, semble avoir un biais nettement moins important que les autres pour $n = 2$, mais non pour les valeurs plus grandes

de n . Pour le cas $n > 2$, nous avons utilisé deux méthodes à

grande entropie, à savoir l'échantillonnage aléatoire à partir d'une population aléatoirement ordonnée (RANSYS) et la

méthode proposée par Tillé (1996) (TILLÉ). Dans tous les estimateurs de la variance, le biais est systématiquement

TILLÉ que pour RANSYS, surtout quand n prend sa valeur la plus grande de 40. Les différences entre les biais de

TILLÉ et ceux de RANSYS sont également positives pour toutes les valeurs de n et, de nouveau, particulièrement

quand $n = 40$. Nous conjecturons que cette méthode de sélection TILLÉ a une entropie un peu plus faible (et, typiquement, une variance plus faible) que la méthode

RANSYS.

REMERCIEMENTS

Les auteurs remercient P.S. Kott d'avoir proposé l'équation (10) lors d'une communication personnelle, ainsi qu'un examinateur anonyme pour trois autres suggestions qui leur ont permis d'améliorer l'article.

BIBLIOGRAPHIE

- ASOK, C., et SUKTHATME, B. V. (1976). On samford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 71, 912-918.
- BREWER, K. R. W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.
- BREWER, K. R. W. (1999). Le calage esthétiquement dans le cas de l'échantillonnage avec probabilités inégales. *Techniques d'enquête*, 25, 231-239.
- COCHRAN, W. G. (1963). *Sampling Techniques*. 2^e Ed. New York : John Wiley & Sons, Inc.
- DEVILLE, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 219-230.
- FURNIVAL, G. M., GREGOIRE, T. G., et GROSENBACH, L. R. (1987). Adjusted inclusion probabilities with 3P sampling. *Forest Science*, 33, 617-631.
- GODAMBE, V. P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society B*, 17, 269-278.
- GODAMBE, V. P., et JOSHI, V. M. (1965). Admissibility and Bayes estimation in sampling finite populations I, II, and III. *Annals of Mathematical Statistics*, 36, 1707-1742.
- HÁJEK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.
- HÁJEK, J. (1981). *Sampling from a finite population*. New York : Marcel Dekker.
- HANSEN, M. H., et HURWITZ, W. N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- HARTLEY, H. O., et RAO, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.
- HORVITZ, D. G., et THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- KISH, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.
- RAO, J. N. K. (1963). On three procedures of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 58, 202-215.
- RAO, J. N. K., et BAYLESS, D. L. (1969). An empirical study of the stability of estimators and variance estimators in unequal probability sampling of two units per stratum. *Journal of the American Statistical Association*, 64, 520-559.
- ROSEN, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62, 159-191.
- SAMPFORD, M. R. (1962). *An Introduction to Sampling Theory*. Edinburgh and London : Oliver and Boyd Ltd.
- SAMPFORD, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.
- SÄRDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.
- SEN, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127.
- SUKTHATME, P. V. (1954). *Sampling Theory of Surveys with Applications*. Ames, Iowa : Iowa State College Press.
- TILLÉ, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika*, 83, 238-241.
- YATES, F. (1981). *Sampling Methods for Census and Surveys*. 4^e Ed. London : Charles Griffin and Co.
- YATES, F., et GRUNDY, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, 15, 235-261.

mentionnés à la section précédente (toutefois, pour RANSYS, nous utilisons l'approximation de Hartley et Rao (1962) des π_{ij} au lieu des π_{ij} exactes dans la formule (2)). Nous répétons ce processus d'échantillonnage-estimation $R=50\,000$ fois.

Le tableau 4 montre les biais relatifs de Monte Carlo des estimateurs de la variance pour RANSYS et TILLÉ. À noter, que pour TILLÉ, aucune valeur n'est fournie à la ligne correspondant à l'estimateur de la variance SYG, parce que, étant donné les populations, les mesures de taille et les tailles d'échantillon employées, la méthode de TILLÉ produit des π_{ij} strictement positives, ce qui signifie que l'estimateur de la variance SYG est sans biais par rapport au plan de sondage. Tous les chiffres du tableau sont relativement faibles, ce qui semble confirmer notre opinion selon laquelle, dans des conditions de grande entropie, le calcul des π_{ij} n'est pas essentiel à l'obtention d'estimateurs de la variance presque sans biais. Dans le groupe des estimateurs indépendants de π_{ij} , nous n'observons aucune différence notable en ce qui concerne RANSYS, mais \hat{V}_{HVI}^{DEV} et son parent, \hat{V}_{HVI}^{DEV} , semblent donner d'un peu meilleurs résultats que la famille d'estimateurs $\hat{V}_{16,*}$, en ce qui concerne TILLÉ, particulièrement pour $n=40$. Cependant, pour TILLÉ, tous les biais observés sont positifs et ont tendance à augmenter parallèlement à la taille d'échantillon. Il semble donc que l'entropie de TILLÉ est un peu plus faible que celle de RANSYS, auquel cas les biais plus importants observés pour la famille d'estimateurs $\hat{V}_{16,*}$ reflètent l'effet réel assez exactement.

Tableau 4

BR (%) des estimateurs de la variance pour $n > 2$

Estimateur de la variance		$n = 10 \quad n = 20 \quad n = 40 \quad n = 10 \quad n = 20 \quad n = 40$									
TILLÉ											
\hat{V}_{SYG}	$\hat{V}_{16,18}$	0,13	1,02	-0,27	-	-	-	-	-	-	-
\hat{V}_{HVI}	\hat{V}_{HVI}	-0,14	0,47	-2,35	1,49	2,18	3,27	-	-	-	-
\hat{V}_{DEV}	\hat{V}_{DEV}	-0,12	0,54	-2,15	1,52	2,25	3,48	-	-	-	-
$\hat{V}_{16,9}$	$\hat{V}_{16,9}$	-0,10	0,83	-0,52	1,58	2,54	5,21	-	-	-	-
$\hat{V}_{16,10}$	$\hat{V}_{16,10}$	-0,23	0,64	-0,75	1,41	2,34	4,97	-	-	-	-
$\hat{V}_{16,11}$	$\hat{V}_{16,11}$	0,11	1,02	-0,30	1,75	2,73	5,45	-	-	-	-
$\hat{V}_{16,18}$	$\hat{V}_{16,18}$	0,13	1,03	-0,29	1,77	2,74	5,45	-	-	-	-
BL220											
\hat{V}_{SYG}	\hat{V}_{SYG}	-0,43	0,77	-0,43	0,64	1,01	1,93	-	-	-	-
\hat{V}_{HVI}	\hat{V}_{HVI}	-0,4	-0,75	-0,59	0,67	1,09	2,14	-	-	-	-
\hat{V}_{DEV}	\hat{V}_{DEV}	-0,37	-0,68	-0,39	0,67	1,09	2,14	-	-	-	-
$\hat{V}_{16,9}$	$\hat{V}_{16,9}$	-0,34	-0,51	0,67	0,70	1,26	3,22	-	-	-	-
$\hat{V}_{16,10}$	$\hat{V}_{16,10}$	-0,40	-0,58	0,58	0,63	1,19	3,13	-	-	-	-
$\hat{V}_{16,11}$	$\hat{V}_{16,11}$	-0,27	-0,43	0,76	0,77	1,34	3,31	-	-	-	-
$\hat{V}_{16,18}$	$\hat{V}_{16,18}$	-0,27	-0,43	0,76	0,78	1,34	3,32	-	-	-	-
MU281											
\hat{V}_{SYG}	\hat{V}_{SYG}	-0,27	-0,43	0,77	-	-	-	-	-	-	-
\hat{V}_{HVI}	\hat{V}_{HVI}	-0,4	-0,75	-0,59	0,64	1,01	1,93	-	-	-	-
\hat{V}_{DEV}	\hat{V}_{DEV}	-0,37	-0,68	-0,39	0,67	1,09	2,14	-	-	-	-
$\hat{V}_{16,9}$	$\hat{V}_{16,9}$	-0,34	-0,51	0,67	0,70	1,26	3,22	-	-	-	-
$\hat{V}_{16,10}$	$\hat{V}_{16,10}$	-0,40	-0,58	0,58	0,63	1,19	3,13	-	-	-	-
$\hat{V}_{16,11}$	$\hat{V}_{16,11}$	-0,27	-0,43	0,76	0,77	1,34	3,31	-	-	-	-
$\hat{V}_{16,18}$	$\hat{V}_{16,18}$	-0,27	-0,43	0,76	0,78	1,34	3,32	-	-	-	-

À fin de vérifier si l'entropie de TILLÉ est un peu plus faible que celle de RANSYS, nous comparons leurs variances de Monte Carlo (VMC) au moyen de la formule (12), qui est l'approximation sous grande entropie de c_p c'est-à-dire (18), pour calculer (12). La comparaison est présentée au tableau 5. Celui-ci montre que les variances TILLÉ sont un peu plus faibles que les variances correspondantes RANSYS. En outre, les variances approximatives données par (12) concordent mieux avec les variances RANSYS. Ces résultats appuient notre conjecture voulant que l'entropie de TILLÉ soit un peu plus faible que celle de RANSYS, particulièrement quand la correction pour sondage dans une population finie est appréciable.

Nous nous concentrons ensuite sur la stabilité. Le tableau 6 donne les erreurs-types de Monte Carlo observées des estimateurs de la variance. De toute évidence, aucune différence méritant d'être mentionnée n'existe entre ces estimateurs. Il en est de même dans le cas d'une comparaison des deux méthodes d'échantillonnage. Il semble que la stabilité ne soit pas un facteur pertinent lorsqu'il s'agit de faire en choix entre ces estimateurs de la variance.

Tableau 5

Comparaison des variances (toutes les valeurs en 10^3)

		BL220									
		58,31	41,16	30,70	57,43	40,41	29,54	57,90	40,49	29,48	29,08
\hat{V}_{SYG}	\hat{V}_{SYG}	58,31	41,16	30,70	57,43	40,41	29,54	57,90	40,49	29,48	29,08
\hat{V}_{HVI}	\hat{V}_{HVI}	57,90	40,49	29,48	57,39	40,24	29,08	57,90	40,49	29,48	29,08
\hat{V}_{DEV}	\hat{V}_{DEV}	57,90	40,49	29,48	57,39	40,24	29,08	57,90	40,49	29,48	29,08
$\hat{V}_{16,9}$	$\hat{V}_{16,9}$	57,90	40,49	29,48	57,39	40,24	29,08	57,90	40,49	29,48	29,08
$\hat{V}_{16,10}$	$\hat{V}_{16,10}$	57,90	40,49	29,48	57,39	40,24	29,08	57,90	40,49	29,48	29,08
$\hat{V}_{16,11}$	$\hat{V}_{16,11}$	58,04	40,64	29,73	57,53	40,39	29,32	58,04	40,64	29,73	57,53
$\hat{V}_{16,18}$	$\hat{V}_{16,18}$	58,05	40,65	29,73	57,55	40,39	29,32	58,05	40,65	29,73	57,55
MU281											
\hat{V}_{SYG}	\hat{V}_{SYG}	54,90	37,29	25,33	55,07	37,50	25,45	54,90	37,29	25,33	55,07
\hat{V}_{HVI}	\hat{V}_{HVI}	54,69	36,98	24,96	54,79	37,07	24,78	54,69	36,98	24,96	54,79
\hat{V}_{DEV}	\hat{V}_{DEV}	54,68	36,98	24,95	54,79	37,07	24,77	54,68	36,98	24,95	54,79
$\hat{V}_{16,9}$	$\hat{V}_{16,9}$	54,67	36,92	24,70	54,77	37,01	24,52	54,67	36,92	24,70	54,77
$\hat{V}_{16,10}$	$\hat{V}_{16,10}$	54,63	36,89	24,66	54,74	36,98	24,48	54,63	36,89	24,66	54,74
$\hat{V}_{16,11}$	$\hat{V}_{16,11}$	54,70	36,95	24,74	54,81	37,04	24,56	54,70	36,95	24,74	54,81
$\hat{V}_{16,18}$	$\hat{V}_{16,18}$	54,71	36,96	24,74	54,81	37,04	24,56	54,71	36,96	24,74	54,81

caractéristiques indépendantes de π_{ij} . Donc, il s'agit de « concurrents » dans la même catégorie.

On peut évaluer les propriétés d'un estimateur de la variance de diverses façons: ici, nous nous concentrerons sur le *biais* et la *stabilité*. Nous présentons les principaux résultats de nos études aux sections qui suivent. Nous considérons séparément deux cas, à savoir $n = 2$ et $n > 2$.

5.1 Cas $n = 2$

En vue de tester les estimateurs de la variance dans diverses situations, nous avons utilisés dans l'étude neuf petites populations dont la plupart étaient également incluses dans les études de stabilité réalisées par Rao et Bayless (1969). Le tableau 1 résume les caractéristiques principales de chaque population, y compris les coefficients de variation (CV) de y et de x , et le coefficient de correction pour la variance de y et de x . Ici, y est la variable pour laquelle nous voulons estimer les totaux et x est une variable auxiliaire qui peut être utilisée pour la sélection de l'échantillon. Notons que N varie de 10 à 20, CV (x), de 0,14 à 0,73, et p , de 0,49 à 0,94. Ceci nous donne un bon mélange de populations ayant des caractéristiques différentes.

Les probabilités d'inclusion sont choisies de sorte qu'elles soient proportionnelles à x_i , c'est-à-dire $\pi_i = 2X_i / X$, pour tout i . Nous considérons aussi deux plans d'échantillonnage, à savoir la procédure de Brewer (1963) (appelée BREWER) et la procédure d'élimination de Tillé (1996) (appelée TILLÉ). Pour les deux méthodes, les π_{ij} sont faciles à calculer et, pour ces neuf populations, elles sont strictement positives (cette condition n'est pas toujours satisfaite par la méthode TILLÉ). En outre, puisque $n = 2$, pour tout échantillon $s = \{i, j\}$, nous avons $p(s) = \pi_{ij}$. Donc, nous pouvons obtenir les propriétés statistiques exactes de tout estimateur donné de la variance V . Pour cela, représentons par S l'ensemble de tous les échantillons possibles de taille $n = 2$ pour une population U . L'espérance de \hat{V} est alors définie comme étant

$$E(\hat{V}) = \sum_{s \in S} p(s) \hat{V}(s),$$

et son erreur-type $(E, -T, \cdot)$, comme étant

$$E, -T, (\hat{V}) = \left\{ \sum_{s \in S} p(s) [\hat{V}(s) - E(\hat{V})]^2 \right\}^{1/2}.$$

Pour chacun des deux plans d'échantillonnage sus-mentionnés, le tableau 2 donne le *biais relatif* $BR(\hat{V}) = E(\hat{V}) / V(X^{HT}) - 1$, exprimé en pourcentage, pour les six estimateurs de la variance indépendants de π_{ij} . Les deux premiers de ces estimateurs ne nécessitent aucune explication; les quatre autres correspondent à (16) conjugué à (9), (10), (11) et (18), respectivement. Puisque pour $n = 2$ (uniquement), \hat{V}^{DEV} et $\hat{V}^{16,9}$ sont identiques, ils figurent tous deux sur la même ligne. Pour simplifier la lecture du tableau, nous avons mis en relief le BR le plus faible (en valeur absolue) pour chaque population et plan d'échantillonnage.

L'examen du tableau 2 suscite les commentaires qui suivent. Premièrement, les propriétés des estimateurs de la variance indépendants de π_{ij} sont assez bonnes pour toutes les populations, sauf, éventuellement, pour la population 4. L'examen de la relation entre x et y pour cette population révèle l'existence d'une certaine courbure, la croissance de la population des grandes villes étant plus rapide. Il existe aussi une valeur aberrante (ville 26) pour laquelle le nombre de personnes a presque triplé dans l'intervalle de dix ans entre 1920 et 1930. Un autre cas intéressant est celui des populations 5 et 6. Ces deux populations ayant des définitions identiques, on s'attendrait à obtenir des résultats comparables. Cependant, les chiffres de BR pour la population 5 sont nettement moins bons que pour la population 6, particulièrement pour la méthode BREWER. La seule différence observable entre ces deux populations est que la population 5 contient une valeur aberrante (ferme 14 dans la référence fournie). Il semblerait donc que l'existence de valeurs aberrantes puisse introduire un certain biais supplémentaire dans ces estimateurs de la variance. Deuxièmement, l'estimateur $\hat{V}^{16,18}$ semble être le meilleur de la catégorie, puisqu'il donne d'assez bons résultats dans toutes les situations et produit les biais les plus faibles (en valeur absolue) dans la plupart des cas. Troisièmement, l'estimateur $\hat{V}^{16,10}$ a tendance à produire les biais les plus importants.

Pour ce qui est de la stabilité, le tableau 3 donne le coefficient de variation $CV(\hat{V}) = E, -T, (\hat{V}) / E(\hat{V})$, exprimé en pourcentage, pour les sept estimateurs de la variance. On constate que les estimateurs de la variance indépendants de π_{ij} ont tendance à être plus efficaces (coefficients de variation plus petits) que \hat{V}^{SYG} , bien que les gains soient faibles. À part cela, peu de choses distinguent ces estimateurs de la variance les uns des autres, même si $\hat{V}^{16,10}$ est celui qui donne les meilleurs résultats pour toutes les populations sauf la dernière.

5.2 Cas $n > 2$

À la présente section, nous adoptons une méthode en simulation de Monte Carlo pour examiner les propriétés de l'estimateur de la variance, T , étudié porte sur deux populations réelles. La première est une population de 220 blocs (BL220) tirée de l'annexe B dans Kish (1965). T , ensemble de données contient deux variables : X_i = nombre de logements occupés par des locataires dans le bloc i et X_i = nombre total de logements dans le bloc i . Certains caractéristiques de cette population sont : CV (y) = 1,06, CV (x) = 0,96 et $p = 0,99$.

La deuxième population, qui comprend 281 municipalités (MU281), est donnée dans Sarnadal, Swensson et Wrethman (1992). Nous utilisons RMT85, c'est-à-dire les recettes de l'imposition de 1985, comme variable étudiée, y , et P75, c'est-à-dire la population municipale en 1975, comme mesure de taille. Les caractéristiques principales de cette population sont :

Une autre propriété séduisante de l'estimateur (16) est la suivante. Quand c_i est spécifiée par (9), nous avons

$$c_i^{-1} - \pi_i = (n - \pi_i)/(n - 1) \{1 - \pi_i\} \quad (17)$$

Le facteur $(1 - \pi_i)$ peut être interprété facilement comme une correction pour sondage dans une population finie, tandis que le facteur $n/(n - 1)$ joue un rôle complètement différent, qu'on peut expliquer comme suit. Il est facile de voir que $\beta = X^{HT} n^{-1}$ est un estimateur sans biais par rapport au modèle de β dans le modèle (13). Posons que $\hat{\sigma}_i^2 = (X_i - \beta \pi_i)^2$ pour tout i . Alors, $(X_i^{HT} n^{-1})^2 = (X_i - \beta \pi_i)^2 \pi_i^2 = \hat{\sigma}_i^2 \pi_i^2$, $i \in U$. Il n'est pas difficile de montrer que le facteur $n/(n - 1)$ élimine le biais (du au modèle) de $\sum_{i \in s} (X_i^{HT} n^{-1})^2 = \sum_{i \in s} \hat{\sigma}_i^2 \pi_i^2$ en tant qu'estimateur de $\sum_{i \in s} \sigma_i^2 \pi_i^2$.

Le choix de (9) pour préciser la valeur de c_i simplifie aussi particulièrement le calcul de l'estimation HT proprement dite et de sa variance estimative; par substitution de (17) dans (16) et développement de cette expression en termes individuels, nous obtenons :

$$\hat{V}(Y^{HT}) = \{n/(n - 1)\} \left\{ \sum_{i \in s} Y_i^2 \pi_i^{-2} - n^{-1} Y_2^{HT} \sum_{i \in s} Y_i \pi_i^{-2} \right\}$$

$$- \sum_{i \in s} Y_i^2 \pi_i^{-1} + 2n^{-1} Y_2^{HT} \sum_{i \in s} Y_i - n^{-2} Y_2^{HT} \sum_{i \in s} \pi_i \}$$

Cette formule comprend six expressions, c'est-à-dire $n Y_2^{HT}$, $\sum_{i \in s} Y_i^2 \pi_i^{-2}$, $\sum_{i \in s} Y_i^2 \pi_i^{-1}$, $\sum_{i \in s} Y_i$, $\sum_{i \in s} \pi_i$ qui sont les sommes d'échantillon de 1 (unité), Y_1^{HT} , Y_2^{HT} , X_1^{HT} et X_2^{HT} , respectivement. Si nous cumuloons ces termes individuels sur chaque unité d'échantillonnage, nous pouvons alors évaluer ensemble Y_2^{HT} et $V(Y^{HT})$ en un seul passage machine des données d'échantillon.

Notons qu'en cas de non-réponse, on peut obtenir une correction de premier ordre en conditionnant l'échantillon sur la taille d'échantillon réalisée, que nous pouvons représenter ici par n' . Cela comprendrait le remplacement des probabilités d'inclusion de premier ordre originales, π_i , par les « probabilités d'inclusion corrigées », $\pi_i' = \pi_i n/n'$. (Cette terminologie est tirée de Fumival, Gregoire et Grosebaugh (1987), qui ont utilisé le même genre de correction dans un contexte différent). Les sommations sur l'échantillon réalisés, s' , deviendraient alors n' , $\sum_{i \in s'} X_i' \pi_i'^{-1}$, $\sum_{i \in s'} X_i'^2 \pi_i'^{-2}$, $\sum_{i \in s'} X_i' Y_i'$, et $\sum_{i \in s'} \pi_i'$, respectivement.

Au-delà des propriétés énumérées plus haut, nous pouvons poursuivre l'étude de (16) à l'aide du modèle ξ donné par (13). L'expression la plus désirable de l'espérance sous ξ d'un estimateur de $V(Y^{HT})$ est $\sum_{i \in s} \sigma_i^2 \pi_i^{-1} (n_i^{-1} - 1)$, car cette expression a, elle-même, l'espérance par rapport au plan de sondage $\sum_{i \in U} \sigma_i^2 \pi_i^{-1}$, qui est la borne inférieure de la variance anticipée de tout estimateur sans biais (Godambe 1955; Godambe et Joshi 1965). Pour chacune des trois définitions de c_i , l'espérance sous ξ de (16) diffère de $\sum_{i \in s} \sigma_i^2 \pi_i^{-1} (n_i^{-1} - 1)$ par

des termes d'ordre $O(N^{-1})$. Bien que ces termes aient tendance à s'annuler, ils ne sont pas entièrement négligeables, puisqu'ils sont uniquement $O(N^{-1})$ plus petits que la variance proprement dite. Compte tenu de cela, nous avons jugé souhaitable d'utiliser une nouvelle version de c_i , retenant les propriétés (plan de sondage) (i) à (iii) pour (16) et fournissant une expression s'approchant davantage de $\sum_{i \in s} \sigma_i^2 \pi_i^{-1} (n_i^{-1} - 1)$ pour l'espérance sous ξ de (16). Ces exigences sont satisfaites par un c_i défini comme suit :

$$c_i = (n - 1) \left\{ n - (2n - 1) (n - 1)^{-1} \pi_i + (n - 1)^{-1} \sum_{k \in U} \pi_k^2 \right\}, \quad (18)$$

pour tout $i \in U$. Avec cette définition de c_i , l'espérance sous ξ de (16) contient encore certains termes « indésirables », mais il ne s'agit maintenant que d'un terme unique d'ordre $O(N^{-2})$, qui est par conséquent inférieur d'un facteur d'ordre $O(N^{-1} n^{-1})$ à $V(Y^{HT})$, et d'autres termes encore plus petits.

5. CERTAINS RÉSULTATS EMPIRIQUES

En vue d'évaluer les propriétés de l'estimateur de la variance proposé à la section 4, nous avons réalisé certaines études empiriques. Nous avons aussi inclus dans ces études trois autres estimateurs de la variance, à savoir i) l'estimateur SYG, donné par (2), ii) l'estimateur de la variance proposé par Hájek (1964, page 1520),

$$\hat{V}_{H\dot{A}J}(Y^{HT}) = \{n/(n - 1)\} \sum_{i \in s} (1 - \pi_i) (X_i^{HT} \pi_i^{-1} - A_s)^2, \quad (19)$$

où $A_s = \sum_{i \in s} a_i X_i^{HT} \pi_i^{-1}$, $a_i = (1 - \pi_i) / \sum_{k \in s} (1 - \pi_k)$ et iii) une légère modification de (19) proposée par Deville (1999),

$$\hat{V}_{DEV}(Y^{HT}) = \frac{1}{2} \sum_{i \in s} \frac{1 - \sum_{k \in s} a_k^2}{(1 - \pi_i) (X_i^{HT} \pi_i^{-1} - A_s)^2}. \quad (20)$$

Il convient de mentionner qu'au départ, l'estimateur (19) a été proposé uniquement pour un plan de sondage à grande entropie particulier, à savoir le tirage avec rejet, et non pour tous les plans de sondage à grande entropie. Cependant on a proposé ultérieurement l'utilisation de cet estimateur pour certains autres plans de sondage à grande entropie. Par exemple, Rosen (1997) a suggéré d'utiliser (19) dans le contexte de l'échantillonnage de Pareto.

L'inclusion des estimateurs (2), (19) et (20) dans nos études empiriques mérite une brève explication. L'estimateur SYG serait normalement le favori si les π_{ij} étaient connues et n'étaient ni nulles ni très petites comparativement aux $\pi_i \pi_j$ correspondantes. Dans ces conditions, il serait naturel de se demander s'il existe une différence significative, en ce qui concerne les propriétés, entre (2) et l'estimateur plus simple (16). Par ailleurs, une comparaison de (19) et (20) est intéressante, car ces deux estimateurs ont en commun avec (16) leur simplicité et les

conséquent, nous utilisons ici l'expression légèrement différente donnée par (11), $(1 - 2n^{-1}\pi_i + n^{-2}\sum_{k \in U} \pi_k^2)$ étant les deux premiers termes du développement en série de Taylor de la valeur inverse de $(1 + 2n^{-1}\pi_i - n^{-2}\sum_{k \in U} \pi_k^2)$ et

inversement.

L'étape suivante consiste à remplacer les π_{ij} dans le

troisième terme de (7) par l'approximation (8). Ce

remplacement donne

$$- \sum_{i \in U} \sum_{j \neq i} (\pi_i \pi_j - \pi_{ij}) (X_i \pi_i^{-1} - X_j \pi_j^{-1} - X_i \pi_i^{-1} - X_j \pi_j^{-1}) \\ = - \sum_{i \in U} \sum_{j \neq i} \pi_i \pi_j [1 - (c_i + c_j)/2]$$

$$(X_i \pi_i^{-1} - X_j \pi_j^{-1})(X_i \pi_i^{-1} - X_j \pi_j^{-1}) \\ = \sum_{i \in U} (1 - c_i) \pi_i^2 (X_i \pi_i^{-1} - X_j \pi_j^{-1})^2,$$

et donc la variance de l'estimateur HT peut être approximée

$$\hat{V}(Y^{\text{HT}})$$

Cette variance approximative a une forme très simple. Elle est aussi sans erreur sous *eassr* pour chacun des trois choix de c_i présentés plus haut.

3. UNE VÉRIFICATION ASSISTÉE PAR MODÈLE DE L'UTILITÉ DE LA FORMULE DE LA VARIANCE APPROXIMATIVE

Considérons le modèle de ratio qui suit comme étant une description possible de la population échantillonnée :

$$\xi_i : Y_i = \beta \pi_i + e_i; E \xi_i = 0; E \xi_i^2 = \sigma_\xi^2;$$

$$E \xi_i (e_j) = 0, i \neq j; i, j \in U. \quad (13)$$

Il s'agit d'un modèle abrégé destiné à refléter la situation où les valeurs prévues Y_i sont *intrinsèquement* proportionnelles aux valeurs X_i d'une variable auxiliaire x et les probabilités d'inclusion π_i sont *choisies* de sorte qu'elles soient proportionnelles aux X_i . Il est naturellement impossible que les Y_i dépendent directement des probabilités d'inclusion proprement dites, puisque ces probabilités peuvent être fixées assez arbitrairement par la personne concevant l'échantillon.

La prédiction ou l'espérance fondée sur le modèle, sous ξ_i , de l'expression de la variance approximative (12) est

4. ESTIMATION DE LA VARIANCE PAR RAPPORT AU PLAN DE SONDAGE DE L'ESTIMATEUR HT

La présente section a pour objectif de proposer un estimateur d'échantillon plausible de la variance par rapport au plan de sondage approximative de l'estimateur HT donne par (12). Un tel estimateur est

$$c = (1 - n^{-1}) \sum_{i \in U} \sigma_i^2 / \sum_{i \in U} \sigma_i^2 \left(1 - 2n^{-1} \pi_i - n^{-2} \sum_{k \in U} \pi_k^2 \right). \quad (15)$$

Sous *eassr*, (15) devient $c = N(n - 1) / \{n(N - 1)\}$, qui donne l'expression exacte pour $\hat{V}(Y^{\text{HT}})$. Même sans *eassr*, le remplacement de σ_i^2 par $\sigma^2 \pi_i$ dans (15) donne (10) pour c . Il est rassurant de constater que l'analyse fondée purement sur le plan de sondage et celle assistée par modèle produisent des résultats aussi concordants.

On e., $\sum_{i \in U} e_i$. Idéalement, l'expression (14) devrait être égale à $E \hat{V}(Y^{\text{HT}})$, à savoir $\sum_{i \in U} \sigma_i^2 (\pi_i^{-1} - 1)$ (Godambe 1955; Godambe et Joshi 1965). Cette condition mène à la formule implicite

$$E \hat{V}(Y^{\text{HT}}) = E \xi_i \sum_{i \in U} \pi_i (1 - c_i \pi_i) (e_i \pi_i^{-1} - e_i \pi_i^{-1})^2 \\ = \sum_{i \in U} \sigma_i^2 \left\{ \pi_i^{-1} - n^{-1} - c_i (1 - 2n^{-1} \pi_i) - n^{-2} \sum_{k \in U} c_k \pi_k^2 \right\} \quad (14)$$

qu'on obtient en remplaçant chaque somme de population (12) par l'estimateur HT correspondant et en corrigeant par le facteur c_i^{-1} . Cet estimateur possède certaines propriétés intéressantes : i) pour chacun des trois c_i choisis, il se réduit à l'estimateur type de la variance dans le cas de l'*eassr*; ii) il est facile à calculer, puisqu'il ne comporte aucune double sommation et iii) par la technique de linéarisation de Taylor, on peut montrer que (16) est approximativement sans biais par rapport au plan de sondage pour (12).

$$\hat{V}(Y^{\text{HT}}) = \sum_{i \in U} (c_i^{-1} - \pi_i) (X_i \pi_i^{-1} - Y_i^{\text{HT}} \pi_i^{-1})^2, \quad (16)$$

deuxième ordre peuvent s'écrire $\pi_{ij} = \pi_i \pi_j$

[$N(n-1)/\{n(N-1)\}$]. Le facteur $N(n-1)/\{n(N-1)\}$ est

inférieur à 1, et tend vers 1 pour les grandes populations et

tailles d'échantillon. Pour ce plan de sondage, le troisième

terme de (7) expéhic seulement $1/N$ de la variance

complète de l'estimateur HT. De surcroît, pour plusieurs

plans d'échantillonnage à probabilité proportionnelle à la

taille, comme l'échantillonnage avec rejet (Hájek 1964) et

l'échantillonnage systématique aléatoire avec probabilité

proportionnelle à la taille (PPT) (Hartley et Rao 1962), la

condition $\pi_{ij} \approx \pi_i \pi_j$ est également vérifiée, à condition que

N et n soient suffisamment grands.

Il existe toutefois certaines exceptions où le troisième

terme de (7) peut être grand. La plus importante de ces

exceptions est le tirage systématique à partir d'une

population dont les unités sont arrangées dans un ordre

significatif avant le tirage. Dans un tel cas, un certain

nombre de probabilités d'inclusion de deuxième ordre

peuvent même être nulles. Cette situation et d'autres cas

spéciaux doivent être examinés séparément et ne font

l'objet d'aucune autre discussion dans le présent article.

Le reste de la présente section est consacré à l'établisse-

ment d'une approximation de $V(\hat{X}^{HT})$ fondée uniquement

sur des probabilités d'inclusion de premier ordre. Nous

commençons par proposer une approximation simple de π_{ij}

de la forme

(8) $\pi_{ij} \approx \pi_i \pi_j (c_i + c_j) / 2, \quad i \neq j \in U.$

Trois choix possibles pour $c_i, i \in U$, sont alors :

(9) $c_i = (n - 1) / (n - \pi_i),$

(10) $c_i = (n - 1) / \left(n - \sum_{j \in U} \pi_j \right) \text{ et}$

(11) $c_i = (n - 1) / \left(n - 2\pi_i + \sum_{j \in U} \pi_j^2 \right).$

Les deux premiers choix de c_i sont guidés par des ratios

de sommes des π_{ij} aux sommes correspondantes des $\pi_i \pi_j$.

Donc, d'une part, nous obtenons la formule (9) en

comparant (3) à (4), et, d'autre part, la formule (10) en

comparant (5) à (6). Enfin, la formule (11) est fondée sur

les expressions asymptotiques de π_{ij} obtenues par Hartley

et Rao (1962) et par Asok et Sukhatne (1976) pour

l'échantillonnage systématique aléatoire πPT et pour la

procédure de Sampford (1967), respectivement. Jusqu'à

l'ordre $O(n^{-3})$, ces expressions asymptotiques se

simplifient toutes deux en

$\pi_{ij} = \pi_i \pi_j \{ (n-1) / n \} \left\{ 1 + n^{-1} (\pi_i + \pi_j) - \sum_{k \in U} \pi_k^2 \right\},$

qui, à son tour, implique $c_i = \{(n-1)/n\}$

de c_i ne donne pas la formule exacte pour les π_{ij} . Par

fixe (voir Hájek 1981), les probabilités d'inclusion de

l'entropie parmi l'ensemble des plans de sondage à taille

aléatoire simple sans remise (*eassr*), qui maximise

être satisfait par les plans d'échantillonnage à grande

entropie. Par exemple, dans le cas de l'échantillonnage

comparativement aux deux autres. Cette condition semble

attendre à un troisième terme de valeur très faible dans (7)

$\pi_{ij} \approx \pi_i \pi_j$, pour tout $i \neq j \in U$, alors nous pouvons

du plan d'échantillonnage $p(s)$. Donc, si $p(s)$ est tel que

La grandeur du troisième terme dépend principalement

important, ni l'un ni l'autre ne dépend des π_{ij} .

plausible de la variance complétée de l'estimateur HT et, fait

deux termes constituent une première approximation

une population finie. Par conséquent, ensemble, ces

être considéré comme une correction pour sondage dans

chaque tirage étant $p_i^* = \pi_i / n, i \in U$. Le deuxième terme peut

avec remise, la probabilité de sélectionner l'unité i lors de

(1943) du total pour un échantillonnage comptant n tirages

variance de l'estimateur correspondant de Hansen-Hurwitz

Le premier terme de (7) est virtuellement le même que la

(10) $\sum_{i \in U} \sum_{j \in U, j \neq i} (\pi_i \pi_j - \pi_{ij}) (X_i X_j - X_i X_j - X_i X_j - X_i X_j)$

$= \sum_{i \in U} \pi_i^2 (X_i^2 - 2X_i X_j + X_j^2) - \sum_{i \in U} \pi_i^2 (X_i^2 - 2X_i X_j + X_j^2)$

$V(\hat{X}^{HT})$

En utilisant les relations (3) et (4), nous pouvons montrer

que l'équation ci-dessus est identique à

(11) $\sum_{i \in U} \sum_{j \in U, j \neq i} (\pi_i \pi_j - \pi_{ij}) (X_i X_j - X_i X_j - X_i X_j - X_i X_j)$

La formule de rechange s'obtient comme suit. Nous

commençons par une modification triviale de (1),

La variance sous grande entropie de l'estimateur de Horvitz-Thompson

K.R.W. BREWER et MARTIN E. DONADIO¹

RÉSUMÉ

Au moyen d'arguments fondés purement sur le plan de sondage d'une part et sur un modèle d'autre part, nous montrons que, dans des conditions de grande entropie, la variance de l'estimateur de Horvitz-Thompson (HT) dépend presque entièrement des probabilités d'inclusion de premier ordre. Nous établissons des expressions approximatives et des estimateurs de cette variance sous « grande entropie » de l'estimateur HT. Nous réalisons des études en simulation de Monte Carlo pour examiner les propriétés statistiques des estimateurs proposés de la variance.

MOTS CLÉS : Estimateur de Horvitz-Thompson; échantillonnage assisté par modèle; simulation de Monte Carlo; estimation de la variance.

1. INTRODUCTION

Soit U une population finie de N unités étiquetées $i = 1, \dots, N$, et soit X_i la valeur de la i^{e} unité d'une certaine caractéristique y . Considérons le problème de l'estimation du total de la population $X_{\cdot} = \sum_{i=1}^N X_i$. Si un échantillon, s , de n unités de la population U est tiré sans remise avec des probabilités d'inclusion de premier ordre $\pi_i, i \in U$, l'estimateur de Horvitz-Thompson (HT) (1952) du total est $X_{\cdot HT} = \sum_{i \in s} X_i / \pi_i$. Dans le présent article, nous nous limitons aux plans d'échantillonnage à taille fixe. Pour ce cas spécial important, Sen (1953), ainsi que Yates et Grundy (1953) ont montré indépendamment que la variance de $X_{\cdot HT}$ est

$$V(X_{\cdot HT}) - (1/2) \sum_{i \in U} \sum_{j \in U, j \neq i} (\pi_i \pi_j - \pi_{ij}) (X_i / \pi_i - X_j / \pi_j)^2, \quad (1)$$

où π_{ij} est la probabilité de deuxième ordre ou probabilité d'inclusion conjointe des i^{e} et j^{e} unités dans l'échantillon. Ils ont par conséquent suggéré l'estimateur de la variance

$$\hat{V}_{\text{SYG}}(X_{\cdot HT}) = (1/2) \sum_{i \in s} \sum_{j \in s, j \neq i} \pi_i^{-1} (\pi_i \pi_j - \pi_{ij}) (X_i / \pi_i - X_j / \pi_j)^2. \quad (2)$$

On sait que cet estimateur donne de meilleurs résultats que Thompson (1952) (ce dernier étant, cependant, habituellement sans biais pour un n aléatoire), mais le fait que (2) dépende de façon critique de π_{ij} s'est avéré problématique (Brewer 1999). Si une ou plusieurs des $N(N-1)/2$ valeurs distinctes de π_{ij} sont nulles, l'estimateur (2) présente un biais par défaut et si n importe laquelle de ces valeurs est très faible comparativement aux valeurs correspondantes de $\pi_i \pi_j$, (2) est instable (autrement dit, il sera lui-même sujet à une forte variance). En outre, la double sommation figurant dans (2) est assez peu commode, particulièrement

pour les échantillons de grande taille. Il existe non seulement un beaucoup plus grand nombre de π_{ij} qu'il n'y a de π_i , mais il est aussi fréquent que les π_{ij} individuelles soient difficiles à évaluer. Étant donné ces difficultés, l'objectif du présent article est de proposer d'autres estimateurs de la variance qui ne dépendent pas des π_{ij} et qui sont faciles à calculer.

À la section suivante, nous présentons une nouvelle expression de la variance par rapport au plan de sondage de l'estimateur HT. Cette nouvelle expression mène, dans des conditions de grande entropie, à l'établissement d'une formule approximative pour $V(X_{\cdot HT})$ ne contenant pas π_{ij} . À la section 3, nous vérifions l'utilité de notre formule de modèle. À la section 4, nous proposons un estimateur de notre variance approximative qui devrait donner de bons résultats dans des conditions de grande entropie (c'est-à-dire l'absence de toute régularité ou ordonnancement décelable dans les unités d'échantillonnage sélectionnées). Cependant, la plupart des scénarios de sélection d'échantillon produisent des échantillons à grande entropie. En vue de tester l'utilité de l'estimateur de la variance présenté à la section 4, nous avons réalisé certaines études empiriques. À la section 5, nous présentons les résultats de ces études et à la section 6, certaines conclusions.

2. CERTAINES FORMULES APPROXIMATIVES POUR LA VARIANCE PAR RAPPORT AU PLAN DE SONDAGE DE L'ESTIMATEUR HT

Nous commençons par présenter une formule de rechange de la variance de l'estimateur HT, valide unique-ment quand le plan d'échantillonnage prévoit une taille fixe d'échantillon. Avant de procéder, nous énonçons les relations suivantes, qui seront utiles plus tard :

¹ Ken Brewer, School of Finance and Applied Statistics, Australian National University, ACT 0200, Australie. Courriel : Ken.Brewer@anu.edu.au et Martin E. Donadio, Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australie. Courriel : M.Donadio@abs.gov.au.

- FIEENBERG, S.E., MAKOV, U.E. et STEELE, R.J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, 14, 485-502.
- FIEENBERG, S.E., STEELE, R.J. et MAKOV, U.E. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: Data swapping and log-linear models. Dans *Proceedings of Bureau of Census 1996 Annual Research Conference*, 87-105.
- FRANCONI, L. et STANDER, J. (2002). A model based method for disclosure limitation of business microdata. *The Statistician*, 51, 1-11.
- FRANCONI, L., et STANDER, J. (2003). Spatial and non-spatial model-based protection procedures for the release of business microdata. *Statistics and Computing*. Forthcoming.
- FULLER, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383-406.
- KENNICKELL, A.B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. Dans *Record Linkage Techniques, 1997*, W. Alvey et B. Jamerson (Eds.). Washington, D.C.: National Academy Press, 248-267.
- LITTLE, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.
- LIU, F., et LITTLE, R.J.A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. Dans *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2133-2138.
- POLETTINI, S. (2003). Maximum entropy simulation for microdata protection. *Statistics and Computing*. À venir.
- POLETTINI, S., FRANCONI, L. et STANDER, J. (2002). Model-based disclosure protection. Dans *Inference Control in Statistical Databases*, J. Domingo-Ferrer (Ed.). Berlin: Springer-Verlag, 83-96.
- RAGHUNATHAN, T.E., REITER, J.P. et RUBIN, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.
- REITER, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18, 531-544.
- REITER, J.P. (2003). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Rapport technique, Institute of Statistics and Decision Sciences, Duke University.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.
- RUBIN, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462-468.
- WILLENBORG, L., et DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

Le tableau 4 montre que, si l'on impute toutes les valeurs des variables, l'augmentation de la valeur de m au-delà de cinq produit des gains d'efficacité appréciables. L'importance du gain dépend de la grandeur de b_m . S'il est petit relativement à v_m , par exemple lorsqu'on impute des valeurs uniquement pour un petit nombre d'unités sélectionnées, le gain d'efficacité dû à l'augmentation de la valeur de m n'est pas important. Pour toute stratégie fondée sur des données partiellement synthétiques, les imputeurs peuvent comparer les gains d'efficacité aux compromis éventuels en ce qui concerne la confidentialité grâce à des études en simulation du comportement des intrus pour divers nombres d'ensembles de données synthétiques diffusés.

Tableau 4
Sensibilité des inférences sur données partiellement synthétiques à la valeur de m

Conditions	Var_{q_m}	Moy. T_p	Couv. IC à 95 %
Inférence pour β			
$m = 2$	6,52	6,5	92,7
$m = 3$	5,38	5,38	94,4
$m = 4$	4,64	4,89	95,4
$m = 5$	4,46	4,54	95,1
$m = 10$	3,87	3,88	94,4
$m = 50$	3,3	3,37	95,1
Inférence pour α			
$m = 2$	10,62	10,89	93,4
$m = 3$	8,92	9,15	94,9
$m = 4$	8,41	8,45	94,9
$m = 5$	7,69	7,94	95,4
$m = 10$	6,99	7,02	94,8
$m = 50$	6,05	6,28	95,5
Inférence pour γ_4			
$m = 2$	8,13	7,96	93,4
$m = 3$	6,51	6,86	95,5
$m = 4$	6,11	6,33	95,6
$m = 5$	5,83	6	95,3
$m = 10$	5,13	5,38	95,4
$m = 50$	4,66	4,87	95,5

Les variances associées à α sont multipliées par 10⁶.

5. CONCLUSION

Les simulations présentées dans l'article montrent que les règles habituelles de combinaison des ensembles de données obtenus par imputation multiple peuvent produire des estimations de la variance présentant un biais par excès quand on les applique à des données partiellement synthétiques. Les nouvelles règles présentées ici semblent corriger le problème, donc produire des inférences plus fiables. D'autres études seront nécessaires pour évaluer les propriétés de ces nouvelles règles lorsqu'on applique une

REMERCIEMENTS

La présente étude a été financée par le United States Bureau of the Census aux termes d'un contrat avec Datametrics Research. L'auteur remercie Trivellore Raghunathan, Donald Rubin et Laura Zayatz de leur appui statistique et de leurs encouragements au cours des travaux, ainsi que l'examineur et un rédacteur adjoint de leurs suggestions et commentaires constructifs.

Enfin, le présent article ne traite pas de l'effet de diverses stratégies fondées sur des données partiellement synthétiques sur la protection de la confidentialité ni de la comparaison des méthodes fondées sur des données partiellement synthétiques à d'autres méthodes de contrôle de la divulgation. Ces comparaisons aideraient les imputeurs à déterminer si les premières méthodes conviennent pour leurs fichiers de microdonnées à grande diffusion.

BIBLIOGRAPHIE

ABOWD, J.M., et WOODOCK, S.D. (2001). Disclosure limitation in longitudinal linked data. Dans *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, EP. Doyle, J. Lane, L. Zayatz et J. Theeuwes (Eds.). Amsterdam: North-Holland. 215-277.

DANDEKAR, R.A., COHEN, M. et KIRKENDAL, N. (2002a). Sensitive micro data protection using Latin hypercube sampling technique. Dans *Inference Control in Statistical Databases*, J. Domingo-Ferrer (Ed.). Berlin: Springer-Verlag. 117-125.

DANDEKAR, R.A., DOMINGO-FERRER, J. et SEBE, F. (2002b). LHS-based hybrid microdata versus rank swapping and microaggregation for numeric microdata protection. Dans *Inference Control in Statistical Databases*, J. Domingo-Ferrer (Ed.). Berlin: Springer-Verlag. 153-162.

loi prédictive a posteriori bayésienne de $(X^4 | q_{\text{obs}})$, obtenue en ajustant la régression de X_4 sur (X^1, X^2, X^3) . Toutes les unités sont telles que $Z^1_i = 1$ et sont utilisées comme sources de données pour les lois a posteriori. Les paramètres à estimer sont les mêmes que ceux décrits à la section 4.1.2. Le tableau 3 résume les résultats de 5 000 exécutions de la simulation au moyen de $m = 5$ ensembles de données partiellement synthétiques. Pour tous les paramètres à estimer, la moyenne de \bar{q}_5 est presque identique à celle de q_{obs} . En outre, les estimations de la variance fondées sur T^p sont proches de la variance réelle de q_5 . Le léger biais par excès est dû au fait que \bar{v}_m a tendance à surestimer v_{obs} comme nous l'avons expliqué à la section 3.1. En moyenne, T^m surestime $\text{Var}(q_5)$ d'un facteur supérieur à deux et T^s sous-estime fortement $\text{Var}(\bar{q}_5)$ pour α et X^4 . Ces problèmes ne sont pas dus au fait que m est petit, puisqu'ils persistent dans les simulations où m est grand. Même si des erreurs de cette importance ne se produisent pas nécessairement dans d'autres conditions, les résultats de cet exemple simple indiquent de nouveau que T^m et T^s ne sont,

en général, pas appropriés pour analyser des données partiellement synthétiques, surtout quand on produit des valeurs entièrement synthétiques pour certaines variables. Les imputeurs ont de bonnes raisons de ne diffuser qu'un petit nombre d'ensembles de données synthétiques. Chaque ensemble de données supplémentaire nécessite un espace de stockage supplémentaire et, par dessus tout, la diffusion d'un trop grand nombre d'ensembles de données risque de compromettre le respect de la confidentialité, car des intrus pourraient combiner les valeurs imputées pour obtenir des renseignements sur les valeurs réelles. Le tableau 4 donne les résultats pour des répétitions indépendantes de 5 000 exécutions de la simulation basées sur diverses valeurs de m . Les estimations ponctuelles sont sans biais pour les trois paramètres estimés et ne sont donc pas présentes dans le tableau. Les taux de couverture des intervalles de confiance à 95 % sont proches de 95 % pour toutes les valeurs de m supérieures à deux. La surestimation dans le cas de T^p est de nouveau due à un biais par excès

Tableau 2

Résultats des simulations où les valeurs de X_4 sont imputées pour les unités pour lesquelles $X^1_i > 1$

Type d'inférence	Paramètre : β	SÉLECTION	TOUT	Données observées*	Couverture des IC à 95 %				
					Moy. \bar{q}_5	Var \bar{q}_5	Moy. T^p	Moy. T^m	En utilisant T^p En utilisant T^m
Paramètre : β	SÉLECTION	10,02	5,45	5,89	5,68	8,97	95,3 %	98,2 %	96,9 %
		TOUT	10,04	5,28	7,57	93,7 %	95,5 %	96,5 %	97,9 %
		Données observées*	10	4,7	95,4 %	94,1 %	95,4 %	96,5 %	97,9 %
		SÉLECTION	$9,25 \times 10^3$	$4,49 \times 10^6$	$4,76 \times 10^6$	$6,97 \times 10^6$	$95,4 \%$	$94,1 \%$	$96,5 \%$
Paramètre : α	SÉLECTION	$9,25 \times 10^3$	$4,49 \times 10^6$	$5,03 \times 10^6$	$4,75 \times 10^6$	$6,31 \times 10^6$	$95,4 \%$	$94,1 \%$	$96,5 \%$
		TOUT	$9,59 \times 10^3$	$4,26 \times 10^6$	$4,26 \times 10^6$	$6,31 \times 10^6$	$95,4 \%$	$94,1 \%$	$96,5 \%$
		Données observées*	$-2,34 \times 10^3$	$4,76$	$5,59$	$93,8 \%$	$94,5 \%$	$96,6 \%$	$95,4 \%$
		TOUT	$-1,24 \times 10^3$	$4,82$	$6,09$	$95,0 \%$	$95,4 \%$	$96,6 \%$	$95,4 \%$
Paramètre : X^4	SÉLECTION	$-1,45 \times 10^2$	$4,97$	$5,01$	$6,09$	$95,0 \%$	$95,4 \%$	$96,6 \%$	$95,4 \%$
		TOUT	$-1,24 \times 10^3$	$4,82$	$5,59$	$93,8 \%$	$94,5 \%$	$96,6 \%$	$95,4 \%$
		Données observées*	$-2,34 \times 10^3$	$4,76$	$5,59$	$93,8 \%$	$94,5 \%$	$96,6 \%$	$95,4 \%$
		TOUT	$-1,24 \times 10^3$	$4,82$	$6,09$	$95,0 \%$	$95,4 \%$	$96,6 \%$	$95,4 \%$

* Les titres de colonne ne s'appliquent pas à cette ligne. Il s'agit de la moyenne de q_{obs} de la variance de q_{obs} et du pourcentage des intervalles de confiance à 95 % pour les données observées qui couvrent Q .

Tableau 3

Résultats des simulations lors de l'imputation de toutes les valeurs d'une variable

Paramètre à estimer	Moy. q_{obs}	Moy. \bar{q}_5	Var q_{obs}	Var \bar{q}_5	Moy. T^p	Moy. T^m	Moy. T^s
β	9,9500	9,9400	3,19	4,46	4,54	11,1	4,63
α	0,0137	0,0135	6,12	7,69	7,94	17,3	5,17
X^4	0,0000	0,0000	4,55	5,83	6,00	12,3	2,87

Si les imputations se fondent sur TOUTES les unités — une méthode d'imputation incorrecte — dans le cas du scénario « aléatoire », T_p présente un biais par défaut et 92,6 % seulement des intervalles de confiance à 95 % synthétiques contiennent la valeur zéro, L'utilisation de T_m fait passer le taux de couverture à 95 %, ce qui donne à penser qu'il est plus sûr d'utiliser T_m que T_p lorsqu'on utilise TOUTES les unités pour l'imputation. Les intervalles de confiance fondés sur la méthode TOU et sur T_m sont, en moyenne, plus grands que ceux fondés sur la méthode SELECTION et sur T_p . Ce résultat illustre l'avantage qu'il y a à conditionner sur Z pour obtenir des imputations correctes, même si le scénario utilisé pour fixer $Z_j = 1$ ne dépend pas des valeurs de X_j .

L'estimateur de la variance pour les données entièrement synthétiques, T_s , qui n'est pas présente au tableau 1, est négatif pour chacune des 5 000 simulations pour les deux scénarios et pour les deux méthodes d'imputation. Manifestement, bien qu'il soit valide pour les données entièrement synthétiques (Raghnathan et coll., 2003), T_s n'est en général pas approprié pour les données partiellement synthétiques.

4.1.2 Simulations au moyen de quatre variables

Chaque ensemble de données observé, D , comprend $n = 200$ valeurs de quatre variables, $(X_1^2, X_2^2, X_3^2, X_4^2)$, générées comme suit : $(Y_1^2, Y_2^2, Y_3^2) \sim MVN(0, \Sigma)$, où Σ est telle que toutes les variances sont égales à 1 unité et que toutes les covariances sont égales à 0,5, et $(Y_4^2 | Y_1^2, Y_2^2, Y_3^2) \sim N(10Y_1^2 + 7Y_2^2 + 4Y_3^2, 25^2)$. Pour fixer les idées, nous pouvons considérer la variable X_1^2 comme étant un identificateur clé et X_4^2 comme étant une variable de nature délicate. Le plan est de simuler des valeurs de la variable de nature délicate X_4^2 pour toutes les unités ayant une valeur « inhabituelle » de l'identificateur clé, définie comme étant $X_1^2 > 1$. Donc, X_j^{imp} comprend les valeurs échantillonnées de (X_1^2, X_2^2, X_3^2) et les valeurs de X_4^2 pour les unités pour lesquelles $X_1^2 \leq 1$. Habituellement, environ 30 unités par ensemble de données observé sont telles que $X_1^2 > 1$.

Comme précédemment, nous examinons deux scénarios pour déterminer la loi prédictive a posteriori pour les imputations. La méthode SELECTION consiste à utiliser uniquement les unités pour lesquelles $Z_j = 1$ comme sources de données pour les lois a posteriori tandis que la méthode TOU consiste à utiliser toutes les unités observées. Sous chaque scénario, pour procéder aux imputations, i) nous tirons des valeurs des paramètres de la régression de X_4^2 sur (X_1^2, X_2^2, X_3^2) à partir de leur loi a posteriori qui est estimée en utilisant les unités SELECTIONNES ou TOUTES les unités et ii) nous tirons des valeurs de X_4^2 pour les unités pour lesquelles $Z_j = 1$ en nous servant des valeurs tirées pour les paramètres. En tout, $m = 5$ ensembles de données synthétiques sont générés pour chaque ensemble de données observé D . Les paramètres d'intérêt incluent β , le coefficient de régression de X_4^2 dans la régression linéaire de X_4^2 sur

Le tableau 2 résume les résultats pour 5 000 exécutions de cette simulation. Quand les imputations sont fondées sur les unités SELECTIONNES, les moyennes de \hat{q} et de T_p s'écartent de moyennes de q^{obs} et $\text{Var}(\hat{q}_j)$ d'une valeur égale ou inférieure à l'erreur de simulation. En outre, les taux de couverture des intervalles de confiance à 95 % synthétiques sont semblables à ceux des intervalles de confiance à 95 % pour les données observées. Les valeurs de T_m sont considérablement plus élevées que celles de $\text{Var}(\hat{q}_j)$, ce qui donne des taux de couverture d'environ 97 %. Les valeurs de T_s , qui ne sont pas présentes au tableau 2, sont négatives pour chacune des 5 000 exécutions de la simulation. Dans l'ensemble, ces résultats confirment ceux présentés à la section 4.1.1 : si les valeurs imputées sont tirées d'une loi a posteriori conditionnée sur Z_j , les estimations ponctuelles et les estimations d'intervalle fondées sur T_p sont plus exactes que celles fondées sur T_m ou sur T_s .

Même si les imputations fondées sur TOUTES les unités ne sont pas correctes, il est instructif d'examiner les propriétés de T_p et de T_m pour ce genre d'imputation. Les imputeurs pourraient fonder leurs imputations sur toutes les unités observées pour des raisons pratiques, par exemple parce que les unités pour lesquelles $Z_j = 1$ ne fournissent pas suffisamment de données pour ajuster les modèles d'imputation. Les résultats reflètent ceux de la section 4.1.1 : T_p sous-estime $\text{Var}(\hat{q}_j)$, ce qui donne des taux de couverture d'environ 94 %, tandis que l'utilisation de T_m porte les taux de couverture à environ 96 %, principalement à cause du biais par excès dans T_m . De nouveau, ces résultats donnent à penser que, si les imputeurs fondent effectivement leurs imputations sur toutes les unités observées bien qu'il n'en existe que quelques-unes pour lesquelles $Z_j = 1$, il est plus prudent que les analystes utilisent T_m au lieu de T_p pour estimer la variance. Comme nous l'avons montré à la section 4.1.1, les intervalles fondés sur TOUTES les unités sont habituellement plus grands que ceux fondés sur les unités SELECTIONNES, si bien que, dans la mesure du possible, les imputeurs devraient fonder leurs imputations uniquement sur les unités pour lesquelles $Z_j = 1$.

4.2 Imputation de toutes les valeurs de X pour une variable

Chaque ensemble de données observé comprend $n = 200$ valeurs de quatre variables générées comme suit : $(Y_1^2, Y_2^2, Y_3^2) \sim MVN(0, \mathbf{I})$, où \mathbf{I} est la matrice d'identité et $(Y_4^2 | Y_1^2, Y_2^2, Y_3^2) \sim N(10Y_1^2 + 10Y_2^2 + 10Y_3^2, 25^2)$. Donc, $X_j^{\text{imp}} = (X_1^2, X_2^2, X_3^2)$. Les valeurs de X_4^2 sont imputées à partir de la

loi prédictive a posteriori bayésienne de $(Y_4^* | Y^{obs})$, obtenue en ajustant la régression de Y_4 sur (X_1, Y_2, Y_3) . Toutes les unités sont telles que $Z_j = 1$ et sont utilisées comme sources de données pour les lois a posteriori. Les paramètres à estimer sont les mêmes que ceux décrits à la section 4.1.2. Le tableau 3 résume les résultats de 5 000 exécutions de la simulation au moyen de $m = 5$ ensembles de données partiellement synthétiques. Pour tous les paramètres à estimer, la moyenne de \bar{q}_5 est presque identique à celle de q^{obs} . En outre, les estimations de la variance fondées sur T_p sont proches de la variance réelle de \bar{q}_5 . Le léger biais par excès est dû au fait que \bar{v}_m a tendance à surestimer v^{obs} , comme nous l'avons expliqué à la section 3.1. En moyenne, $T_m^{suresstime}$ Var(\bar{q}_5) d'un facteur supérieur à deux et T_s sous-estime fortement Var(\bar{q}_5) pour α et Y_4 . Ces problèmes ne sont pas dus au fait que m est petit, puisqu'ils persistent dans les simulations où m est grand. Même si des erreurs de cette importance ne se produisent pas nécessairement dans d'autres conditions, les résultats de cet exemple simple indiquent de nouveau que T_m et T_s ne sont,

dans \bar{v}_m . Les imputeurs ont de bonnes raisons de ne diffuser qu'un petit nombre d'ensembles de données synthétiques. Chaque ensemble de données supplémentaire nécessite un espace de stockage supplémentaire et, par dessus tout, la diffusion d'un trop grand nombre d'ensembles de données risque de compromettre le respect de la confidentialité, car des intrus pourraient combiner les valeurs imputées pour obtenir des renseignements sur les valeurs réelles. Le tableau 4 donne les résultats pour des répétitions indépendantes de 5 000 exécutions de la simulation basées sur diverses valeurs de m . Les estimations ponctuelles sont sans biais pour les trois paramètres estimés et ne sont donc pas présentées dans le tableau. Les taux de couverture des intervalles de confiance à 95 % sont proches de 95 % pour toutes les valeurs de m supérieures à deux. La surestimation dans le cas de T_p est de nouveau due à un biais par excès

Tableau 2

Résultats des simulations où les valeurs de Y_4 sont imputées pour les unités pour lesquelles $Y_1 > 1$

Type d'inférence	Moy. \bar{q}_5	Var \bar{q}_5	Moy. T_p	Moy. T_m	En utilisant T_p	En utilisant T_m	Couverture des IC à 95 %
Paramètre : β							
SÉLECTION	10,02	5,45	5,68	8,97	93,7 %	95,3 %	98,2 %
TOUT	10,04	5,89	5,28	7,57	93,7 %	95,5 %	96,9 %
Données observées*	10	4,7					95,5 %
Paramètre : α							
SÉLECTION	$9,25 \times 10^{-3}$	$4,49 \times 10^{-6}$	$4,76 \times 10^{-6}$	$6,97 \times 10^{-6}$	95,4 %	95,4 %	97,9 %
TOUT	$9,59 \times 10^{-3}$	$5,03 \times 10^{-6}$	$4,75 \times 10^{-6}$	$6,31 \times 10^{-6}$	94,1 %	94,1 %	96,5 %
Données observées*	$9,66 \times 10^{-3}$	$4,26 \times 10^{-6}$					95,4 %
Paramètre : Y_4							
SÉLECTION	$-1,45 \times 10^{-2}$	4,97	5,01	6,09	95,0 %	95,0 %	96,6 %
TOUT	$-1,24 \times 10^{-3}$	5,19	4,82	5,59	93,8 %	93,8 %	95,4 %
Données observées*	$-2,34 \times 10^{-3}$	4,76					94,5 %

* Les titres de colonne ne s'appliquent pas à cette ligne. Il s'agit de la moyenne de q^{obs} , de la variance de q^{obs} et du pourcentage des intervalles de confiance à 95 % pour les données observées qui couvrent \bar{Q} .

Tableau 3

Résultats des simulations lors de l'imputation de toutes les valeurs d'une variable

Paramètre à estimer	Moy. q^{obs}	Moy. \bar{q}_5	Var q^{obs}	Var \bar{q}_5	Moy. T_p	Moy. T_m	Moy. T_s
β	9,9500	9,9400	3,19	4,46	4,54	11,1	4,63
α	0,0137	0,0135	6,12	7,69	7,94	17,3	5,17
Y_4	0,0000	0,0000	4,55	5,83	6,00	12,3	2,87

Si les imputations se fondent sur TOUTES les unités — une méthode d'imputation incorrecte — dans le cas du scénario « aléatoire », T_p présente un biais par défaut et 92,6 % seulement des intervalles de confiance à 95 % synthétiques contiennent la valeur zéro. L'utilisation de T_m fait passer le taux de couverture à 95 %, ce qui donne à penser qu'il est plus sûr d'utiliser T_m que T_p lorsqu'on utilise TOUTES les unités pour l'imputation. Les intervalles de confiance fondés sur la méthode TOUT et sur T_m sont, en moyenne, plus grands que ceux fondés sur la méthode SELECTION et sur T_p . Ce résultat illustre l'avantage qu'il y a à conditionner sur Z pour obtenir des imputations correctes, même si le scénario utilisé pour fixer $Z_j = 1$ ne dépend pas des valeurs de X_j .

L'estimateur de la variance pour les données entièrement synthétiques, T_p , qui n'est pas présenté au tableau 1, est négatif pour chacune des 5 000 simulations pour les deux scénarios et pour les deux méthodes d'imputation. Mani-festement, bien qu'il soit valide pour les données entièrement synthétiques (Raghnunathan et coll. 2003), T_p n'est en général pas approprié pour les données partiellement synthétiques.

4.1.2 Simulations au moyen de quatre variables

Chaque ensemble de données observé, D , comprend $n = 200$ valeurs de quatre variables, $(X_1^1, X_2^1, X_3^1, X_4^1)$, générées comme suit : $(Y_1^1, Y_2^1, Y_3^1) \sim MVN(0, \Sigma)$, où Σ est telle que toutes les variances sont égales à 1 unité et que toutes les covariances sont égales à 0,5, et $(Y_4^1 | Y_1^1, Y_2^1, Y_3^1) \sim N(10Y_1^1 + 7Y_2^1 + 4Y_3^1, 25)$. Pour fixer les idées, nous pouvons considérer la variable X_1^1 comme étant un identificateur clé et X_4^1 comme étant une variable de nature délicate. Le plan est de simuler des valeurs de la variable de nature délicate X_4^1 pour toutes les unités ayant une valeur « inhabituelle » de l'identificateur clé, défini comme étant $X_1^1 > 1$. Donc, X_4^1 comprend les valeurs échantillonnées de (X_1^1, X_2^1, X_3^1) et les valeurs de X_4^1 pour les unités pour lesquelles $X_1^1 \leq 1$. Habituellement, environ 30 unités par ensemble de données observé sont telles que $X_1^1 > 1$.

Comme précédemment, nous examinons deux scénarios pour déterminer la loi prédictive à posteriori pour les imputations. La méthode SELECTION consiste à utiliser uniquement les unités pour lesquelles $Z_j = 1$ comme sources de données pour les lois à posteriori tandis que la méthode TOUT consiste à utiliser toutes les unités observées. Sous chaque scénario, pour procéder aux imputations, i) nous tirons des valeurs des paramètres de la régression de X_4^1 sur (X_1^1, X_2^1, X_3^1) à partir de leur loi a posteriori qui est estimée en utilisant les unités TOUTES ou TOUTES les unités et ii) nous tirons des valeurs de X_4^1 pour les unités pour lesquelles $Z_j = 1$ en nous servant des valeurs tirées pour les paramètres. En tout, $m = 5$ ensembles de données synthétiques sont générés pour chaque ensemble de données observé D . Les paramètres d'intérêt incluent β , le coefficient de régression de X_4^1 dans la régression linéaire de X_4^1 sur

Le tableau 2 résume les résultats pour 5 000 exécutions de cette simulation. Quand les imputations sont fondées sur les unités SELECTIONNES, les moyennes de \hat{q}_j et de T_p s'écartent des moyennes de q_{obs} et $\text{Var}(\hat{q}_j)$ d'une valeur égale ou inférieure à l'erreur de simulation. En outre, les taux de couverture des intervalles de confiance à 95 % synthétiques sont semblables à ceux des intervalles de confiance à 95 % pour les données observées. Les valeurs de T_m sont considérablement plus élevées que celles de $\text{Var}(\hat{q}_j)$, ce qui donne des taux de couverture d'environ 97 %. Les valeurs de T_p , qui ne sont pas présentées au tableau 2, sont négatives pour chacune des 5 000 exécutions de la simulation. Dans l'ensemble, ces résultats confirment ceux présentés à la section 4.1.1 : si les valeurs imputées sont tirées d'une loi à posteriori conditionnée sur Z_j , les estimations ponctuelles et les estimations d'intervalle fondées sur T_p sont plus exactes que celles fondées sur T_m ou sur T_p .

Même si les imputations fondées sur TOUTES les unités ne sont pas correctes, il est instructif d'examiner les propriétés de T_p et de T_m pour ce genre d'imputation. Les imputeurs pourraient fonder leurs imputations sur toutes les unités observées pour des raisons pratiques, par exemple parce que les unités pour lesquelles $Z_j = 1$ ne fournissent pas suffisamment de données pour ajuster les modèles d'imputation. Les résultats reflètent ceux de la section 4.1.1 : T_p sous-estime $\text{Var}(\hat{q}_j)$, ce qui donne des taux de couverture d'environ 94 %, tandis que l'utilisation de T_m porte les taux de couverture à environ 96 %, principalement à cause du biais par excès dans T_m . De nouveau, ces résultats donnent à penser que, si les imputeurs fondaient effectivement leurs imputations sur toutes les unités observées bien qu'il n'en existe que quelques-unes pour lesquelles $Z_j = 1$, il est plus prudent que les analystes utilisent T_m au lieu de T_p pour estimer la variance. Comme nous l'avons montré à la section 4.1.1, les intervalles fondés sur TOUTES les unités sont habituellement plus grands que ceux fondés sur les unités SELECTIONNES, si bien que, dans la mesure du possible, les imputeurs devraient fonder leurs imputations uniquement sur les unités pour lesquelles $Z_j = 1$.

4.2 Imputation de toutes les valeurs de X pour une variable

Chaque ensemble de données observé comprend $n = 200$ valeurs de quatre variables générées comme suit : $(Y_1^1, Y_2^1, Y_3^1) \sim MVN(0, \mathbf{I})$, où \mathbf{I} est la matrice d'identité et $(Y_4^1 | Y_1^1, Y_2^1, Y_3^1) \sim N(10Y_1^1 + 10Y_2^1 + 10Y_3^1, 25)$. Donc, $X_4^1 = (X_1^1, X_2^1, X_3^1)$. Les valeurs de X_4^1 sont imputées à partir de la

$Y \sim N(0, 10^2)$. Nous utilisons deux scénarios distincts pour spécifier les unités pour lesquelles $Z_j = 1$, afin de générer deux ensembles de données partiellement synthétiques pour chaque D . Le premier scénario, appelé « aléatoire », consiste à remplacer X pour 20 unités échantillonnées aléatoirement à partir de D . Le deuxième, appelé « grand Y » consiste à remplacer X uniquement pour les unités pour lesquelles $Y_j > 10$.

Pour chaque D , et pour chaque scénario, il existe $m = 5$ ensembles de données synthétiques, $d_i = (X_{rep,i}^*, I, Z)$, pour $i = 1, \dots, 5$. Les $X_{rep,i}^*$ sont générés selon une technique bootstrap bayésienne (Rubin 1987, pages 123-124) qui consiste à tirer des valeurs de X à partir d'un groupe donneur de valeurs sélectionnées de Y^{obs} . Soit $Y^{elig,j}$ le vecteur de dimensions $n_0 \times 1$ des valeurs de Y^{obs} qui constituent le groupe donneur. Soit $n_{rep}^* = \sum_{j=1}^{100} Z_j^*$. Le procédé bootstrap bayésien se déroule comme suit :

1. Tirer ($n_0 - 1$) nombres aléatoires uniformes. Les classer par ordre ascendant. Appliquer à ces nombres ordonnés l'échantillonnage $a_0 = 0, a_1, a_2, \dots, a_{n_0-1}, a_0 = 1$.
2. Tirer n_{rep} nombres aléatoires uniformes, $n_1, n_2, \dots, n_j, \dots, n_{n_{rep}}$. Pour chacun de ces n_j , imputer $Y^{elig,j}$ quand $a_{j-1} < n \leq a_j$.

Il est peu probable qu'on utilise ce bootstrap bayésien pour imputer des données dans des conditions réelles, puisque les ensembles de données contiennent plus d'une variable. Nous l'employons ici parce qu'il produit des imputations simples, appropriées pour la présente illustration.

Comme nous l'avons mentionné à la section 2, la loi prédictive a posteriori correcte est $f(Y|D, Z)$ et non $f(Y|D)$. Par conséquent, le groupe donneur, Y^{elig} , devrait être égal à l'ensemble $\{X_j : Z_j = 1\}$, que nous nommons « SELECTION ». Aux fins de comparaison, nous imputons

Tableau 1
Résultats des simulations pour l'imputation des valeurs d'une seule variable

Scénario et méthode d'imputation												
Couverture des IC à 95 %												
$Z_j = 1$ pour 20 unités sélectionnées aléatoirement												
$Z_j = 1$ pour les unités pour lesquelles $Y_j > 10$												
SELECTION			0,024	1,097	1,067	1,42	94,5 %	96,7 %	TOUIT			
TOUIT			0,020	1,233	1,044	1,281	92,6 %	94,9 %	SELECTION			
SELECTION			0,016	1,031	1,011	1,068	94,5 %	95,0 %	TOUIT			
TOUIT			-2,383	0,796	0,736	0,921	20,7 %	28,8 %	Résultats pour les données observées*			
			0,016	1,021	1,000		94,5 %					

* Les titres de colonne ne s'appliquent pas à cette ligne. La moyenne de $q^{obs} = 0,016$, la variance de $q^{obs} = 1,021$, la moyenne de $v^{obs} = 1,000$ et 94,5 % des 5 000 intervalles de confiance à 95 % pour les données observées contiennent la valeur zéro.

Le tableau 1 résume les résultats de 5 000 exécutions de cette simulation. Aussi bien pour le scénario « aléatoire » que pour le scénario « grand Y », les moyennes de \bar{q}_j fondées sur les unités SELECTIONNEES sont approximativement égales à la moyenne de q^{obs} . Dans le cas du scénario aléatoire, \bar{q}_j fondé sur TOUITES les unités est également sans biais, car $E(Y_{nrep}^{inap} | X, I) = q^{obs}$ lorsque la moyenne est calculée sur Z (qui est en fait stochastique dans ce scénario). Cependant, si l'on utilise TOUITES les unités dans le scénario « grand Y », q_j présente un biais par défaut important, parce que les valeurs imputées ne sont pas contraintes d'être supérieures à 10 lorsqu'on utilise TOUITES les unités.

Aussi bien dans le scénario « aléatoire » que le scénario « grand Y », 94,5 % des 5 000 intervalles de confiance à 95 % synthétiques fondés sur T^p et les unités SELECTIONNEES contiennent la valeur zéro. Ce taux est identique au taux de couverture de 94,5 % des intervalles de confiance fondés sur les données observées ($q^{obs} \pm 1,96\sqrt{v^{obs}}$). Les taux nominaux sont inférieurs à 95 % à cause de l'erreur de simulation. Les 2 à 3 % d'écart entre les moyennes de T^p et de $\text{Var}(\bar{q}_j)$ équivalent à peu près à l'écart entre les moyennes de v^{obs} et de $\text{Var}(q^{obs})$. L'estimateur habituel de la variance sous imputation multiple, T^m , a tendance à surestimer $\text{Var}(\bar{q}_j)$, ce qui produit un taux de couverture des intervalles de confiance exagérément prudent et montre que T^m n'est pas l'estimateur approprié de la variance lorsqu'on analyse des données partiellement synthétiques correctement imputées.

aussi des valeurs synthétiques au moyen de l'ensemble de donneurs $\{X_j : I_j = 1\}$ que nous nommons « TOUIT ». Les imputations fondées sur TOUITES les unités ne satisfont pas la condition C1 de la section 3.2, puisque $E(q_j | X, Y, I, Z) = (\sum_{j=1}^{100-n_{rep}} Y_{nrep,j}^{inap} + n_{rep} \bar{Y}^{obs}) / n \neq \bar{Y}^{obs}$ tandis que celles fondées sur les unités SELECTIONNEES sont correctes.

Si l'on suppose que les lois a priori de q^{obs} et v^{obs} sont uniformes, la théorie bayésienne classique implique que $(q^{obs}|d_m, B) \sim N(\bar{q}_m, B/m)$ et $(v^{obs}|d_m, B) \sim (v^{obs} < B/m)$. Donc, les moyenne et variance a posteriori de $(\bar{Q}|d_m, B)$ sont

$$E(\bar{Q}|d_m, B) = E(E(\bar{Q}|d_m, B)|d_m, B)$$

$$= E(q^{obs}|d_m, B) = \bar{q}_m \quad (6)$$

$$\text{Var}(\bar{Q}|d_m, B) = E(\text{Var}(\bar{Q}|d_m, B)|d_m, B)$$

$$+ \text{Var}(E(\bar{Q}|d_m, B)|d_m, B)$$

$$= \bar{v}_m + B/m. \quad (7)$$

Puisque toutes les convolutions portent sur des lois normales, $f(\bar{Q}|d_m, B) \sim N(\bar{q}_m, \bar{v}_m + B/m)$.

Pour intégrer cette loi sur $B|d_m$, nous utilisons le fait que $((m-1)b_m^{m-1}|d_m) \sim \chi_{2m-1}^2$ et, suivant l'approximation de Rubin (1987), nous ajustons les deux premiers moments de $\bar{v}_m + B/m$ sur la moyenne quadratique d'une variable aléatoire. L'approximation résultante de la loi a posteriori de \bar{Q} est $(\bar{Q}|d_m) \sim t_{\bar{v}_p}(\bar{q}_m, T_p)$, où \bar{v}_p est telle que définie à la section 2.

3.2 Validité de la randomisation

Pour que les inférences fondées sur les équations (1) à (4) aient des propriétés fréquentistes valides, nous devons imposer deux conditions. Premièrement, l'analyste doit utiliser des estimateurs q et v valides du point de vue de la randomisation. Autrement dit, lorsqu'on applique q et v à D pour obtenir q^{obs} et v^{obs} , $(q^{obs}, v^{obs}|X, Y) \sim N(\bar{Q}, U)$ et $(v^{obs}|X, Y) \sim (U, < < U)$, où la loi pertinente est celle de L . Deuxièmement, les méthodes de génération de données synthétiques doivent être correctes au sens de Rubin (1987). Plus précisément, les méthodes de génération des données doivent satisfaire les conditions suivantes :

C1 : Si l'on fait la moyenne sur les imputations de $X^{rep,i}$, il est nécessaire que

$$\begin{aligned} (i) & (q_i|X, Y, I, Z) \sim N(q^{obs}, B), \\ (ii) & (b_m^{im}|X, Y, I, Z) \sim (B, < < B) \text{ et} \\ (iii) & (v_m^{im}|X, Y, I, Z) \sim (v^{obs}, < < B/m), \text{ où} \end{aligned}$$

C2 : Si l'on fait la moyenne sur les mécanismes d'échantillonnage et de remplacement $(I, Z|X, Y)$, il est nécessaire que $(B|X, Y) \sim (B_0 < < U)$ où $B_0 = E(b_m^{im}|X, Y)$.

Essentiellement, ces conditions exigent que les données synthétiques soient générées de sorte que les q_i soient sans biais pour q^{obs} , b_m^{im} soit sans biais pour B_0 et v_m^{im} soit sans biais pour v^{obs} . Une discussion plus approfondie du procédé d'imputation approprié figure dans Rubin (1987).

Partant de ces hypothèses, il suit que

$$E(\bar{q}_m|X, Y) = E(E(\bar{q}_m|X, Y, I, Z)|X, Y)$$

$$= E(q^{obs}|X, Y) = \bar{Q} \quad (8)$$

$$\text{Var}(\bar{q}_m|X, Y) = E(\text{Var}(\bar{q}_m|X, Y, I, Z)|X, Y)$$

$$+ \text{Var}(E(\bar{q}_m|X, Y, I, Z)|X, Y)$$

$$= E(B|X, Y)/m + \text{Var}(q^{obs}|X, Y) = B_0/m + U. \quad (9)$$

Puisque nous supposons que $(q^{obs}|X, Y)$ et $(q_i^{im}|X, Y, I, Z)$ suivent une loi normale, il s'ensuit que $(\bar{q}_m|X, Y) \sim N(\bar{Q}, B_0/m + U)$.

Si C1 et C2 sont vérifiées, T_p est un estimateur sans biais de $B_0/m + U$. L'approximation de t est justifiée par la méthode de décrit dans Rubin (1987). Plus précisément, l'approximation de t découle du fait que $((m-1)b_m^{m-1}|X, Y) \sim \chi_{2m-1}^2$ et que le nombre de degrés de liberté d'une variable aléatoire suivant la loi du chi carré est égal au double du carré de son espérance sur sa variance.

4. ETUDES EN SIMULATION

La présente section illustre la validité des nouvelles règles de combinaison sous simulation, ainsi que l'efficacité de T_m et de T_s en tant qu'estimateurs de la variance grâce à des études en simulation de stratégies partiellement synthétiques. La section 4.1 décrit deux études où l'imputeur génère des données synthétiques uniquement pour certaines unités. La section 4.2 décrit une étude où l'imputeur génère des données synthétiques pour toutes les valeurs d'une variable d'enquête, en laissant les autres variables à leurs valeurs observées. Les simulations sont basées ici sur des données artificielles et des lois d'imputation correctes pour les imputations. Naturellement, dans les conditions réelles, l'imputeur ne connaît habituellement pas le modèle d'imputation correct et doit l'estimer en s'appuyant sur des données observées et l'expertise disponible du domaine spécialisé. Pour toutes les simulations, les tailles de population sont considérées comme étant infinies afin de pouvoir omettre les facteurs de correction pour population finie.

4.1 Imputation pour certaines unités

L'imputeur peut décider de remplacer les valeurs observées pour quelques unités parmi les données recueillies, puis de diffuser un mélange de valeurs imputées et observées. Nous utilisons cette stratégie dans deux simulations simplistes quoiqu'illustratives, la première comportant une variable et la seconde, quatre variables.

4.1.1 Simulations à l'aide d'une seule variable

Chaque ensemble de données observé, D , comprend $n = 100$ valeurs tirées aléatoirement à partir de

estimateur ponctuel q et la variance de q au moyen d'un estimateur v . Nous supposons que l'analyste détermine les en fait des données recueillies à partir d'un échantillon aléatoire de (X, Y) d'après le plan de sondage réel utilisé pour générer I .

Pour $i = 1, \dots, m$, soit q_i et v_i les valeurs respectives de q et v dans l'ensemble de données synthétiques d_i . Sous certaines conditions, que nous décrivons à la section 3, l'analyste peut obtenir des inférences valides pour la grandeur scalaire \bar{Q} en combinant les q_i et v_i . Précisément, les quantités qui suivent sont nécessaires pour les inférences :

$$(1) \quad \bar{q}_m = \sum_{i=1}^m q_i / m$$

$$(2) \quad b_m = \sum_{i=1}^m (q_i - \bar{q}_m)^2 / (m - 1)$$

$$(3) \quad \bar{v}_m = \sum_{i=1}^m v_i / m$$

L'analyste peut alors utiliser \bar{q}_m pour estimer \bar{Q} et $T^p = b^m / m + \bar{v}^m$ (4)

pour estimer la variance de \bar{q}_m . Si q est une fonction de $(X, Y)_{\text{resp}}$, l'unique est une fonction d'aucune valeur imputée, les inférences à partir des données observées sont identiques aux inférences à partir des données observées; autrement dit, $q_i = q_{\text{obs}}$ et $v_i = v_{\text{obs}}$ pour tout i , et $b^m = 0$. Quand n est grand, les inférences concernant la grandeur scalaire \bar{Q} peuvent se fonder sur des lois de Student à $v^p = (m - 1) / (1 + r_{-1}^m)$ degrés de liberté, où $r^m = (m^{-1} b^m / \bar{v}^m)$. Dans de nombreux cas, r^m et, donc, v^p est suffisamment grand pour qu'une loi normale représentative une approximation adéquate de la loi de Student. Nous ne présentons pas ici les extensions au cas où \bar{Q} est multivarié. T^p diffère de l'estimateur de la variance pour l'imputation multiple de données manquantes, $T^m = (1 + 1/m) b^m + \bar{v}^m$ (Rubin 1987). Dans le contexte de données partiellement synthétiques, \bar{v}^m estime $\text{Var}(q_{\text{obs}})$ et b^m / m estime la variance supplémentaire due à l'utilisation d'un nombre fini d'imputations. Dans le contexte de données manquantes, \bar{v}^m et b^m / m ont la même interprétation, mais un b^m supplémentaire est nécessaire pour obtenir la moyenne sur le mécanisme de non-réponse (Rubin 1987, chapitre 4). Cette moyenne supplémentaire n'est pas nécessaire dans le cas de données partiellement synthétiques, puisque le mécanisme de sélection Z , qui est déterminé par l'imputeur, n'est pas traité comme étant stochastique.

3. JUSTIFICATION DES NOUVELLES RÈGLES DE COMBINAISON

La présente section illustre un calcul bayésien des inférences décrites à la section 2 et les conditions sous lesquelles ces inférences sont valides du point de vue fréquentiste. Ces résultats sont fondés sur la théorie élaborée par Raghunathan et coll. (2003) et la suite de près.

3.1 Calcul bayésien

Pour ce calcul, nous supposons que l'analyste et l'imputeur utilisent le même modèle bayésien. Nous pouvons décomposer la loi a posteriori de $(\bar{Q} | d_m)$, où $d_m = \{d_1, d_2, \dots, d_m\}$, de la façon suivante

$$f(\bar{Q} | d_m) = \int f(\bar{Q} | d_m, D, B) f(D | d_m, B) f(B | d_m) dD dB \quad (5)$$

où $B = \text{Var}(q_i | D, Z)$. En ce qui concerne $f(D | d_m, B)$, l'intégration se fait uniquement sur les valeurs de X_{obs} qui sont remplacées par des valeurs imputées; les composantes de D restent fixes. Sachant D , les données synthétiques sont sans pertinence, de sorte que $f(\bar{Q} | d_m, D, B) = f(\bar{Q} | D)$. Nous supposons que les hypothèses asymptotiques bayésiennes sont vérifiées, si bien que $f(\bar{Q} | D) \sim N(q_{\text{obs}}, v_{\text{obs}})$, où q_{obs} et v_{obs} sont les moyennes et variances a posteriori de \bar{Q} déterminées en utilisant D .

L'intégration de (5) sur D donne $f(\bar{Q} | d_m, B)$. Puisque seules q_{obs} et v_{obs} sont nécessaires pour les inférences au sujet de $(\bar{Q} | D)$, pour $f(D | d_m, B)$, il est suffisant de déterminer $f(q_{\text{obs}} | d_m, B)$. Nous supposons que les imputations sont faites de telle façon que, pour tout i , $(q_i | D, B) \sim N(q_{\text{obs}}, v_{\text{obs}})$ et $(v_i | D, B) \sim (v_{\text{obs}})^{>><}$ notation $F \sim (G, >>< (H))$ signifie que la variable aléatoire F est nettement plus faibles que pour H . En réalité, v_i est habituellement centrée à une valeur supérieure à v_{obs} , puisque les données synthétiques imputaient une incertitude due au tirage des valeurs des paramètres. Pour les échantillons dont la taille n est grande, ce biais devrait être minimal. L'hypothèse selon laquelle $E(q_i | D, B) = q_{\text{obs}}$ devrait être raisonnable si les imputations sont tirées de la loi a posteriori correcte de X pour les unités pour lesquelles $Z_i = 1$.

unités. Toutes ces approches partiellement synthétiques sont intéressantes, car elles promettent d'offrir nombre d'avantages des approches entièrement synthétiques, comme assurer la confidentialité des données en permettant aux utilisateurs de faire des inférences sans être obligés de maintenir des méthodes ou des logiciels statistiques compliqués, tout en réduisant la sensibilité à la spécification des modèles d'imputation.

Bien que des ensembles de données partiellement synthétiques fassent l'objet de grande diffusion, la littérature n'offre pas de renseignements techniques sur la façon de produire des inférences à partir de ces ensembles. À première vue, il peut sembler correct d'utiliser les méthodes d'inférence proposées par Rubin (1987) en cas d'imputation multiple pour remplacer des données manquantes. Malheureusement, comme nous le montrons dans le présent article, ces méthodes produisent parfois des estimations biaisées de la variance. Qui plus est, comme nous le montrons aussi, les méthodes élaborées par Raghunathan et coll. (2003) pour l'analyse des données entièrement synthétiques ne sont pas valides lorsqu'appliquées à des données partiellement synthétiques.

Le présent article décrit les méthodes d'inférence à partir d'ensembles de données partiellement synthétiques obtenus par imputation multiple. L'élaboration de ces méthodes fournit aussi des instructions pour la production de données partiellement synthétiques. La présentation de l'article est la suivante. La section 2 décrit les nouvelles méthodes d'inférence. La section 3 montre l'élaboration de ces méthodes dans un cadre bayésien et décrit les conditions dans lesquelles les inférences résultantes devraient être valides du point de vue fréquentiste. La section 4 décrit les études en simulation qui illustrent la validité de ces méthodes, ainsi que l'inefficacité des règles concurrentes de combinaison de plusieurs estimations ponctuelles et estimations de la variance. La section 5 contient les conclusions et des suggestions quant aux futurs travaux de recherche.

2. INFÉRENCE À PARTIR D'ENSEMBLES DE DONNÉES PARTIELLEMENT SYNTHÉTIQUES OBTENUES PAR IMPUTATION MULTIPLE

Soit $I_j = 1$ si l'on sélectionne l'unité j dans l'enquête originale et $I_j = 0$ autrement. Soit $I = (I_1, \dots, I_N)$. Soit X_j^{obs} la matrice de dimensions $n \times p$ de données recueillies (réelles) pour les unités pour lesquelles $I_j = 1$; soit X_j^{noobs} la matrice de dimensions $(N - n) \times p$ de données d'enquête non observées pour les unités pour lesquelles $I_j = 0$; et soit $X = (X_j^{\text{obs}}, X_j^{\text{noobs}})$. Par souci de simplicité, nous supposons que toutes les unités échantillonnées répondent complètement à l'enquête. Soit X la matrice de dimensions $N \times p$ de variables de plan de sondage pour l'ensemble des N unités de la population, comme des indicateurs de strate

L'organisme qui diffuse des données synthétiques, appelé dans la suite de l'article l'imputeur, construit des ensembles de données synthétiques d'après les données observées, $D = (X_j^{\text{obs}}, Y_j^{\text{obs}})$, selon un processus en deux volets. En premier lieu, il sélectionne parmi les données observées les valeurs qui seront remplacées par des données imputées. En deuxième lieu, il impute les nouvelles valeurs sélectionnées l'unité j pour le remplacement de certaines de ses valeurs observées par des valeurs synthétiques et soit $Z_j = 0$ pour les unités pour lesquelles aucune donnée n'est modifiée. Soit $Z = (Z_1, \dots, Z_n)$. Soit $X_j^{\text{rep}, i}$ toutes les valeurs imputées (remplacées) dans le i^{e} ensemble de données synthétiques et soit $X_j^{\text{rep}, i}$ l'ensemble des valeurs non modifiées (non remplacées) de X_j^{obs} . Nous supposons que les $X_j^{\text{rep}, i}$ sont générées à partir de la loi prédictive a posteriori bayésienne de $(X_j^{\text{rep}, i} | D, Z)$. Les valeurs comprises dans $X_j^{\text{rep}, i}$ sont les mêmes pour tous les ensembles de données synthétiques. Alors, chaque ensemble de données synthétique, d_i , comprend $(X_j, X_j^{\text{rep}, i}, Y_j^{\text{rep}, i}, Z_j)$. Les imputations sont réalisées indépendamment $i = 1, \dots, m$ fois pour produire m ensembles distincts de données synthétiques. Enfin, ces ensembles sont diffusés au public.

Les valeurs contenues dans Z peuvent, et il en est fréquentement ainsi, dépendre des valeurs contenues dans D . Par exemple, l'imputeur peut choisir de ne simuler des variables ou des identificateurs de nature délicate que pour les unités de l'échantillon présentant une combinaison rare d'identificateurs; ou bien, il peut ne remplacer que les valeurs de revenu supérieures à 100 000 \$ par des valeurs imputées. Pour éviter d'introduire un biais, l'imputeur devrait tenir compte de ce genre de sélection en procédant à l'imputation à partir de la loi prédictive a posteriori de X le faire en utilisant uniquement les unités pour lesquelles $Z_j = 1$. En pratique, il peut le faire en utilisant uniquement les unités pour lesquelles $Z_j = 1$ comme source de données pour rechercher les lois a posteriori pour l'imputation. L'utilisation de toutes les unités pour lesquelles $I_j = 1$ peut produire des estimations biaisées ou des intervalles de confiance plus larges avec des taux de couverture exagérément prudents, comme l'illustrent les simulations présentées à la section 4.

À partir de ces ensembles de données synthétiques, l'utilisateur des données diffusées au grand public, appelé dans la suite de l'article l'analyste, veut faire des inférences au sujet d'un paramètre à estimer $\bar{Q} = \bar{Q}(X, Y)$, où la notation $\bar{Q}(X, Y)$ signifie que \bar{Q} est une fonction de (X, Y) . Par exemple, \bar{Q} pourrait représenter la moyenne de population de X ou les coefficients de régression de population de X sur Y . Pour chaque ensemble de données synthétiques d_i , l'analyste estime \bar{Q} au moyen d'un

Inference pour les ensembles de microdonnées à grande diffusion partiellement synthétiques

J.P. REITER¹

RÉSUMÉ

L'une des méthodes permettant d'éviter les divulgations consiste à diffuser des ensembles de microdonnées à grande diffusion partiellement synthétiques. Ces ensembles comprennent les unités enquêtées au départ, mais certaines valeurs recueillies, comme celles de nature délicate présentant un haut risque de divulgation ou celles d'identificateurs clés, sont remplacées par des imputations multiples. Bien qu'on recoure à l'heure actuelle à des approches partiellement synthétiques pour protéger les données à grande diffusion, on ne les a pas encore assorties de méthodes d'inférence valides. Le présent article décrit de telles méthodes. Elles sont fondées sur les concepts de l'imputation multiple en vue de remplacer des données manquantes, mais s'appuient sur des règles différentes pour combiner les estimations ponctuelles et les estimations de la variance. Ces règles de combinaison diffèrent aussi de celles élaborées par Raghunathan, Reiter et Rubin (2003) pour les ensembles de données entièrement synthétiques. La validité de ces nouvelles règles est illustrée au moyen d'études par simulation.

MOTS CLÉS : Confidentialité; divulgation; imputation multiple; données synthétiques.

1. INTRODUCTION

Lors de la diffusion de données au grand public, les organismes statistiques s'efforcent de fournir des données détaillées sans divulguer les renseignements de nature délicate fournis par les répondants. Pour réduire le risque de divulgation, ils modifient habituellement les données originales avant leur diffusion, par exemple, en recodant les variables, en perturbant certaines données ou en ajoutant un bruit aléatoire aux valeurs des variables (Willenborg et de Waal 2001). Cependant, ces méthodes peuvent fausser les liens entre les variables incluses dans l'ensemble de données. Elles compliquent aussi l'analyse car, pour analyser correctement des données perturbées, les utilisateurs devraient suivre les méthodes fondées sur la vraisemblance décrites par Little (1993) ou les modèles d'erreur de mesure décrits par Fuller (1993). Or, ces techniques sont difficiles à appliquer à l'estimation de paramètres non standards et obligeant parfois les analystes à maîtriser de nouvelles méthodes statistiques et de nouveaux logiciels spécialisés.

Rubin (1993) a proposé une autre approche qui consiste à : diffuser des ensembles de données entièrement synthétiques complètement constitués de valeurs produites par imputation multiple au lieu de valeurs réelles. Cette méthode permet de protéger les renseignements confidentiels, puisque l'identification des unités et de leurs données de nature délicate devient difficile quand les données diffuses ne sont pas les valeurs réelles recueillies. En outre, s'ils appliquent les méthodes d'imputation et d'estimation appropriées fondées sur les concepts de l'imputation multiple (Rubin 1987), les utilisateurs des données peuvent obtenir des inférences valides grâce à des

méthodes et des logiciels statistiques standards, applicables aux données complètes. Ces inférences peuvent être obtenues par les méthodes de Raghunathan et coll. (2003), qui reposent sur des règles de combinaison des estimations ponctuelles et des estimations de la variance différentes de celles formulées par Rubin (1987). D'autres discussions et variantes des méthodes fondées sur des données synthétiques figurent dans Little (1993), Fienberg, Steele et Makov (1996), Fienberg, Makov et Steele (1998), Dandekar, Kirkenadall (2002a), Dandekar, Domingo-Ferrer et Sebe (2002b), Poletini et Stander (2002, 2003), Poletini (2003) et Reiter (2002, 2003). Aucun fournisseur de données n'a encore utilisé l'approche entièrement synthétique au stade de la production, mais certains ont adopté une variante, à savoir la diffusion d'ensembles de données partiellement synthétiques comprenant un mélange de valeurs réelles et de valeurs obtenues par imputation multiple. Par exemple, pour protéger les données de la U.S. Survey of Consumer Finances, le U.S. Federal Reserve Board remplace les valeurs monétaires présentant un risque élevé de divulgation par des imputations multiples, puis diffuse un mélange de ces valeurs imputées et de valeurs recueillies non remplacées (Kennickel 1997). Abowd et Woodcock (2001) ont adopté une autre approche partiellement synthétique pour protéger des ensembles de données longitudinales couplées. Ils remplacent toutes les valeurs de certaines variables de nature délicate par des imputations multiples, mais laissent les valeurs d'autres variables telles que l'âge, le sexe, le statut marital, etc., inchangés. Une troisième approche est appliquée par Liu et Little (2002), qui développent un algorithme pour simuler plusieurs valeurs d'identificateurs clés pour certaines données.

$$E[I(f_f = 3)(\gamma_{i2}^2 - \gamma_{j2}^2)]$$

$$= \sum_{f_j=1}^k \sum_{\ell=1}^k \sum_{m=1}^m \pi_{(j\ell)} \pi_{(j\ell m)} \left[\prod_{F_j} (1 - \pi_{(jm)}) \right] \beta_{(j\ell)} \beta_{(jm)}$$

$$= \sum_{f_j=1}^k \sum_{\ell=1}^k \sum_{m=1}^m \pi_{(j\ell)} \pi_{(j\ell m)} \left[\prod_{F_j} (1 - \pi_{(jm)}) \right]$$

$$= (F_f - 1)(F_f - 2) \Pr(f_f = 1),$$

en utilisant la notation de l'annexe 1. Nous pouvons aussi montrer que

$$E[I(f_f = 2)\gamma_{ij}] = (F_f - 1) \Pr(f_f = 1) \quad (\text{A.3})$$

en suivant la preuve de (3) à l'annexe 1, mais en omettant la sommation sur j . (Notons que les membres de (3) sont égaux aux membres correspondants de (A.3) sommés sur j .)

Donc, un estimateur sans biais de $(F_f - 1)^2 \Pr(f_f = 1)$ est

BIBLIOGRAPHIE

- Il s'ensuit que le numérateur de l'expression de $\hat{v}/\hat{\theta}^2$ dans (4) est sans biais pour la deuxième partie de l'expression dans le deuxième membre de (A.2) (en omettant (μ_1/μ_3^2) , ce qu'il fallait démontrer.
- BETHLENHEM, J.G., KELLER, W.J. et PANNKOEK, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.
- HINKINS, S., OH, H.L. et SCHEUREN, F. (1997) Algorithmes de plan de sondage inverses. *Techniques d'enquête*, 23, 13-24.
- SKINNER, C.J., et ELLIOT, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B*, 64, 855-867.
- SKINNER, C.J., et HOLMES, D.J. (1998). Estimating the re-identification risk per record for microdata. *Journal of Official Statistics*, 14, 361-372.
- WILLENBORG, L., et DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. New York : Springer.

qui sous-tend la définition de la mesure pour l'ensemble de la population, θ , à savoir que l'unité de population d'identité connue est sélectionnée aléatoirement à partir de U avec probabilités égales. Une estimation distincte de la mesure dans diverses strates caractérisées par des fractions d'échantillonnage différentes donne aussi une méthode simple de traiter le cas de la sélection avec probabilités inégales. Nous montrons dans le présent article comment tenir compte dans θ et ψ de sources plus générales d'échantillonnage avec probabilités inégales. Les travaux devront se poursuivre afin d'évaluer la robustesse de ces estimateurs lorsqu'on s'écarte de l'hypothèse de l'échantillonnage de Poisson, particulièrement dans le cas de l'échantillonnage à plusieurs degrés.

Lors de l'estimation distincte de la mesure pour diverses sous-populations, l'effet de la réduction de la taille de l'échantillon pourrait être problématique. SE constatent que θ est stable dans leurs études numériques, le coefficient de variation n'excédant jamais 6 %. Toutefois, la taille minimale d'échantillon qu'ils ont étudiée était d'environ 9 000, si bien que d'autres travaux numériques sont nécessaires pour évaluer la stabilité de θ pour de plus petites tailles d'échantillon. La méthode proposée d'estimation de la variance fournit certaines lignes directrices pour tout cas particulier. En principe, on pourrait améliorer la stabilité de l'estimateur en utilisant des hypothèses de modèle et l'un de nous (CJS) poursuit l'étude pour le cas limitant d'une petite sous-population, ne comptant qu'une seule unité, qui étend θ à une mesure du risque au niveau de l'enregistrement analogue à celle considérée par Skinner et Holmes (1998).

ANNEXE 1

Preuve de l'équation (3)

Soit $\beta_i = \pi_i^{j-1} - 1$ et $U_j = \{i \in U; X_i = j\}$, $j = 1, \dots, J$, où X_i représente la valeur de X pour l'unité i . La taille de U_j est F_j . Au lieu d'étiqueter les unités dans U au moyen du simple indice i , considérons le double indice (jk) , $j = 1, \dots, J$, $k = 1, \dots, F_j$, si bien que, dans notre exemple, $\pi_{(jk)}$ représente la probabilité d'inclusion de la k^e unité dans U_j et $\beta_{(jk)} = \pi_{(jk)}^{j-1} - 1$. Sous échantillonnage de Poisson, le deuxième membre de (3) est

$$E \left[\sum_{j=1}^J (F_j - 1) I(F_j = 1) \right] = \sum_{j=1}^J (F_j - 1) \sum_{k=1}^{F_j} \pi_{(jk)} \prod_{l=1}^{l \neq k} (1 - \pi_{(jl)}) \quad (\text{A.1})$$

et le premier membre de (3) est

$$\begin{aligned} &= (\mu_1 / \mu_2^2) \sum_{j=1}^J \left[(F_j - 1)^2 \Pr(f_j = 1) + E\{\gamma_{ij}^2 I(f_j = 2)\} \right] \quad (\text{A.2}) \\ &\approx \text{var} \left[(\mu_1 / \mu_2^2) \sum_{j=1}^J \{(F_j - 1) I(f_j = 1) - \gamma_{ij} I(f_j = 2)\} \right] \\ &\quad \text{var}(\theta - \theta) \end{aligned}$$

Soit $\mu_t = E(\tau_t)$, $t = 1, 2, 3$, et notons que $\mu_1 + \mu_2 = \mu_3$ d'après (3). Une expression linéarisée de $\theta - \theta$ est $\mu_1(-\tau_1 - \tau_2 - \tau_3)/\mu_2^2$, dont la variance peut être exprimée sous la forme

$$\tau_1 = \sum_{j=1}^J I(f_j = 1), \tau_2 = \sum_{j=1}^J I(f_j = 2), \tau_3 = \sum_{j=1}^J F_j I(f_j = 1).$$

Ecrivons $\theta - \theta = \tau_1/(\tau_1 + \tau_2) - \tau_1/\tau_3$, où

Calcul de l'estimateur de la variance par linéarisation

ANNEXE 2

qui est identique à (A.1) et donc (3) s'ensuit.

$$\begin{aligned} &= \sum_{j=1}^J \sum_{k=1}^{F_j} \sum_{l=1}^{F_j} \pi_{(jk)} \pi_{(jl)} \prod_{m=1}^{m \neq k, l} (1 - \pi_{(jm)}) \\ &= \sum_{j=1}^J \sum_{k=1}^{F_j} \sum_{l=1}^{F_j} \pi_{(jk)} \pi_{(jl)} \prod_{m=1}^{m \neq k, l} (1 - \pi_{(jm)}) \\ &= \sum_{j=1}^J \sum_{k=1}^{F_j} \sum_{l=1}^{F_j} \pi_{(jk)} \pi_{(jl)} \prod_{m=1}^{m \neq k, l} (1 - \pi_{(jm)}) \end{aligned}$$

Ceci généralise l'expression de la variance dans la proposition 3 de SE. Nous obtenons l'expression de ψ dans (4) en remplaçant les termes de (A.2) par leurs estimateurs non biaisés. Premièrement, μ_1 et μ_3 sont estimés par τ_1 et τ_3 , respectivement, de sorte que μ_1/μ_3 est estimé par $\theta/(\tau_1 + \tau_2)$. Puis, notons

L'hypothèse de l'échantillonnage de Poisson généralise l'échantillonnage aléatoire simple, l'échantillonnage systématique (avec probabilités égales) ou l'échantillonnage aléatoire stratifié proportionnel. Nous soutenons que, de façon comparable, θ restera approximativement sans biais pour θ sous les plans d'échantillonnage avec probabilités inégales correspondants, c'est-à-dire l'échantillonnage aléatoire simple stratifié disproportionné et l'échantillonnage systématique avec probabilités inégales. Nous soutenons aussi qu'il pourrait être raisonnable de tenir compte de la non-réponse dans θ si s est l'ensemble de répondants et que π_i représente un poids qui pourrait être interprété comme étant la réciproque de la probabilité estimée d'être à la fois échantillonné et répondant.

Comme le mentionnent SF, en pratique, la forme d'échantillonnage qui semble donner lieu au biais le plus important dans θ en tant qu'estimateur de θ est l'échantillonnage à plusieurs degrés, où les unités à plusieurs degrés sont fortement corrélées pour X . Par exemple, le biais pourrait être non négligeable si les ménages forment des grappes dans lesquelles tous les adultes sont échantillonnés, dans le cas où les microdonnées incluent des enregistrements au niveau de l'individu, mais que X est déterminé principalement d'après les variables au niveau du ménage. Ces conditions pourraient donner lieu à une valeur de $n_{(2)}/n_{(1)}$ plus élevée que celle attendue sous échantillonnage de Poisson, donc à une sous-estimation de θ . Toutefois, ce genre d'exemple est quelque peu artificiel et nous pensons que le biais de θ en tant qu'estimateur de θ est modéré dans la plupart des enquêtes sociales types.

4. ESTIMATION DE LA VARIANCE

SF présentent un estimateur par linéarisation de $\text{var}(\theta - \theta)$ qui dépend de $n_{(1)}$ et $n_{(2)}$, comme θ , ainsi que de $n_{(3)} = \sum_{i=1}^J I(f_i = 3)$, le nombre de valeurs de X pour lesquelles il existe exactement trois enregistrements de microdonnées. Nous montrons à l'annexe 2 que, dans le cas de l'échantillonnage de Poisson avec probabilités inégales, on peut généraliser cet estimateur de la variance à

$$\hat{\theta} = \theta_2 \frac{\sum_{i=1}^J \left\{ I(f_i = 3)(\gamma_i^2 - \gamma_i^2) + I(f_i = 2)(\gamma_i^2 + \gamma_i^2) \right\}}{n_{(1)} + \sum_{i=1}^J I(f_i = 2)\gamma_i^2} \quad (4)$$

où $\gamma_{ij} = \sum_s \beta_s^i \gamma_{sj}^i$, $\beta_s^i = \pi_i^{-1} - 1$ et $s_j = \{i \in s; X_i = j\}$, où X_i est la valeur de X pour l'unité i .

5. CONCLUSION

La mesure estimée θ considérée dans le présent article peut servir d'indice pour déterminer si le risque de divulgation associé à un fichier proposé de microdonnées est d'un niveau acceptable ou non. L'objectif peut être de s'assurer que la valeur de θ n'excède pas une probabilité spécifiée. Pour tenir compte de la variation d'échantillonnage dans θ , une méthode plus prudente consisterait à exiger que la borne supérieure d'un intervalle de confiance pour θ , disons $\theta + 2\hat{v}^{1/2}$, n'excède pas la probabilité spécifiée.

En outre, θ peut être utilisé pour comparer diverses stratégies de contrôle du risque de divulgation. Par exemple, on peut inclure dans le fichier de microdonnées des variables présentant des niveaux plus ou moins élevés de détails de classification. Un niveau de détail élevé peut réduire le risque de divulgation si la variable peut être utilisée pour l'appariement à des données externes. On pourrait donc se servir de la mesure estimée θ pour évaluer le risque relatif résultant de diverses méthodes d'agrégation du niveau de classification dans des variables d'identification partielles, y compris la géographie.

On peut estimer la mesure non seulement pour la population dans son ensemble, mais aussi pour des sous-populations. Ce genre de ventilation permet d'obtenir une évaluation plus réaliste du risque que posent les intrus ciblés de cette sorte rend invalide l'hypothèse fondamentale

$$\hat{\theta} = n_{(1)} / \left[n_{(1)} + \sum_{i=1}^J I(f_i = 2)\gamma_i^2 \right].$$

Notons que, selon cette notation, nous pouvons écrire

spécifications de X) et pour une forme de diffusion choisie de sorte que la valeur inférée de θ soit suffisamment faible. Habituellement, il faut procéder à une analyse de sensibilité où la spécification de X varie en fonction non seulement de la forme de diffusion, mais aussi de diverses formes plausibles d'information externe qu'un intrus pourrait posséder au sujet des unités de population connues. Par exemple, on pourrait envisager à la fois un intrus n'ayant accès qu'à de l'information publiquement disponible, et un intrus ayant accès à une base de données privée créée par un organisme.

3. ESTIMATION DE θ

Nous supposons que les données sont les valeurs de X pour les unités d'échantillon f_j , mais non les fréquences des fréquences d'échantillon f_j , mais non les fréquences de la population f_j ($j = 1, \dots, J$). Le « paramètre » d'intérêt, θ , est également inconnu mais doit être estimé. Nous adoptons une méthode d'inférence fondée sur le plan de sondage dans laquelle les f_j sont aléatoires et les F_j sont fixes. Comme l'ont exposé SE, le « paramètre », θ , dépend donc de s , contrairement aux paramètres types de population finie considérés en échantillonnage.

Répéter les étapes suivantes K fois.

Etape 1 : supprimer une unité i de l'échantillon de microdonnées s avec la probabilité

$$a_i = \pi_i / \sum_s \pi_i,$$

où π_i est la probabilité d'inclusion (de premier ordre) de l'unité i ;

Etape 2 : recopier l'unité éliminée dans l'échantillon avec la probabilité π_i ;

Etape 3 : noter si l'unité éliminée correspond à un enregistrement unique dans les microdonnées et si cet appariement est correct.

L'idée est que l'étape 1 imite la sélection par l'intrus (avec probabilité égale) d'une unité à partir de U (en s'appuyant sur la notion de l'échantillonnage inverse de Hinkins. Oh et Scheuren 1997). L'étape 2 imite l'inclusion de cette unité dans s . L'estimateur de θ est la proportion empirique d'appariements uniques qui sont corrects. Suivant l'argument de SE, quand $K \rightarrow \infty$, cet estimateur converge presque certainement vers

$$\theta = \frac{\sum_{s^{(w)}} \Pr(\text{unité } i \text{ éliminée, puis recopiée})}{\sum_{s^{(w)}} \Pr(\text{unité } i \text{ éliminée, puis non recopiée})} + \sum_{s^{(z)}} \Pr(\text{unité } i \text{ éliminée, puis non recopiée})$$

$$= \sum_{s^{(w)}} a_i \pi_i / \left[\sum_{s^{(w)}} a_i \pi_i + \sum_{s^{(w)}} a_i (1 - \pi_i) \right]$$

$$= n^{(1)} / \left[n^{(1)} + \sum_{s^{(z)}} (\pi_i - 1) \right].$$

$$\theta = n^{(1)} / \left[n^{(1)} + \sum_{s^{(z)}} (F_j - 1) I(F_j = 1) \right]. \quad (2)$$

Nous voulons calculer θ , défini par (1), comme étant l'estimateur de θ . SE montrent que θ converge vers θ dans étapes fondamentales de leur argument peuvent être généralisées au cas de l'échantillonnage avec probabilités inégales comme suit. Nous pouvons écrire

$$E \left[\sum_{s^{(w)}} (\pi_i - 1) I(F_j = 1) \right] = E \left[\sum_{s^{(w)}} (F_j - 1) I(F_j = 1) \right]. \quad (3)$$

sous l'hypothèse de l'échantillonnage de Poisson, c'est-à-dire quand les unités de population sont échantillonnées indépendamment. L'équation (3) généralise la proposition 2 de SE. Dans le cas de l'échantillonnage avec probabilités égales, SE montrent comment on peut étendre le résultat de l'équation (3) pour prouver la convergence de θ en tant qu'estimateur de θ , au moyen d'un cadre asymptotique où $J \rightarrow \infty$ et sous certaines conditions de régularité, en particulier le fait que les F_j sont bornées.

Après avoir établi le principal résultat d'absence de biais dans (3), nous conjecturons que ce résultat de convergence peut être généralisé au cas de l'échantillonnage de Poisson avec probabilités inégales, sous réserve de contraintes faibles supplémentaires sur les π_i , par exemple que les π_i aient une borne supérieure constante positive.

Estimation d'une mesure du risque de divulgation pour les microdonnées d'enquête sous échantillonnage avec probabilités inégales

C.J. SKINNER et R.G. CARTER¹

RÉSUMÉ

Skinner et Elliot (2002) ont proposé une mesure simple du risque de divulgation pour les microdonnées d'enquête et montré comment estimer cette mesure sous échantillonnage avec probabilités égales. Dans le présent article, nous montrons comment on peut étendre leurs résultats pour l'estimation de la variance à l'échantillonnage de Poisson et faisons certains commentaires sur les résultats éventuels lorsqu'on s'écarte de cette hypothèse.

MOTS CLÉS : Protection de la confidentialité; inférence en population finie; échantillonnage de Poisson; contrôle de la divulgation statistique; unicité.

1. INTRODUCTION

Les fichiers de microdonnées d'enquête peuvent avoir une grande valeur analytique pour les chercheurs. Lorsqu'ils décident s'ils peuvent donner accès à ce genre de

fichier et de quelle façon, les organismes qui réalisent des enquêtes doivent s'assurer que les données sont protégées contre la divulgation statistique (Willenborg et de Waal 2001). Skinner et Elliot (2002, nommés dans la suite SE) ont proposé une mesure simple du risque de divulgation

statistique relatif aux microdonnées d'enquête qu'on peut utiliser pour prendre des décisions éclairées. Ils ont montré que sous échantillonnage avec probabilités égales, l'estimation de cette mesure est simple. Dans le présent article, nous montrons comment on peut étendre leurs résultats à l'échantillonnage avec probabilités inégales.

À la section 2, nous présentons la mesure. Aux sections 3 et 4, nous envisageons l'estimation ponctuelle et l'estimation de la variance de la mesure, respectivement. Pour le lien entre la mesure proposée et les données publiées sur le risque de divulgation statistique, consulter SE.

2. MESURE DU RISQUE DE DIVULGATION

Nous considérons la diffusion éventuelle d'un fichier de microdonnées comprenant un ensemble d'enregistrements pour les unités (par exemple, des individus ou des ménages) d'un échantillon, sélectionné par une méthode probabiliste à partir d'une population U . Chaque enregistrement correspond à un vecteur de valeurs d'un ensemble spécifique de variables pour l'unité donnée. Suivant une méthode type d'évaluation du risque de divulgation (par exemple, Bethlehem, Keller et Pannekoeck 1990), nous supposons qu'un intrus essaye d'appartier les enregistrements de

où f_j et F_j sont les fréquences des unités de s et de U , respectivement, pour lesquelles $X = j$ et où $I(\cdot)$ est la fonction indicateur ($I(A) = 1$ si A est vrai et $I(A) = 0$, autrement). Le numérateur de θ est le nombre d'enregistrements uniques dans l'ensemble de microdonnées en ce qui concerne X et le dénominateur de θ est le nombre d'unités n importe lequel de ces enregistrements.

La quantité θ est la mesure du risque de divulgation envisagée dans le présent article. Pour assurer la protection contre la divulgation, on doit estimer θ pour diverses formes de diffusion des microdonnées (ce qui sous-entend diverses

$$\theta = \frac{\sum_{j=1}^J I(f_j = 1)}{\sum_{j=1}^J f_j I(f_j = 1)},$$

$\theta = \Pr(\text{appariement correct} | \text{appariement unique})$

L'affirmation soit correcte est :

Nous supposons que l'intrus est capable de déterminer la valeur de J est très grande.)

Les combinaisons possibles de leurs valeurs définissent les catégories $1, \dots, J$ d'une variable X . Habituellement, la valeur de J est très grande.)

Nous supposons que l'intrus est capable de déterminer la valeur de X pour une unité de population d'identité connue et qu'il « affirme » qu'un enregistrement de microdonnées est identifié si, et uniquement si, cette valeur concorde avec la valeur de X figurant dans le fichier de microdonnées pour une seule enregistrement. Si nous supposons a) que l'unité de population d'identité connue est sélectionnée au hasard à partir de U avec probabilités égales et b) que la valeur de X pour cette unité est mesurée de la même façon que X est mesurée dans les microdonnées, la probabilité que l'affirmation soit correcte est :

microdonnées à des unités connues de population à l'aide d'un sous-ensemble spécifique de variables. Nous supposons que ces « variables d'identification » sont nominales et que les combinaisons possibles de leurs valeurs définissent les catégories $1, \dots, J$ d'une variable X . Habituellement, la

¹ C.J. Skinner, University of Southampton, Southampton, United Kingdom, SO17 1BJ et R.G. Carter, Statistique Canada, B-2 Immeuble Jean Talon, Ottawa (Ontario) K1A 0T6.

BIBLIOGRAPHIE

DI PIETRO, E. (1999). Anagrafe informatizzata e Censimenti demografici: dal censimento tradizionale al censimento basato sugli Archivi. *Società Italiana di Statistica: Atti Del Convegno "Verso i Censimenti del 2000*. Udine 7-9 giugno. 169-182.

FORTINI, M. (1994). Un'applicazione del modello a classi latenti per l'analisi dell'errore di copertura del XIII censimento della popolazione. *Atti della XXXVII Riunione Scientifica della Società Italiana di Statistica*. San Remo 6-8 Aprile. 2, 423-430.

GELMAN, A., et RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequence. *Statistical Science*. 7, 457-72.

LAWLESS, J.F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*. 15, 209-225.

MOURA, F.A.S., et HOLT, D. (1999). Production d'estimations régionales à partir de modèles multivariés. *Techniques d'enquête*. 25, 81-89.

SCHWARTZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics*. 6, 461-464.

SPIEGELHALTER, D.J., THOMAS, A., BEST, N. et GILKS, W.R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling*. Version 0.50. Rapport Technique, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

THERNEAU, T.M., et ATKINSON, E.J. (1997). *An Introduction to Recursive Partitioning Using the RPART Routines*. Rapport Technique, Mayo Foundation.

WOLTER, K. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*. 81, 338-346.

ABBATE, C., MASSIELLO, M. et SIGNORE M. (1993). A combined post-enumeration survey for the 1991 Italian population and industrial censuses. *Bulletin of the International Statistical Institute, Firenze, 48th Session*. Tome LV, 2, 159-173.

ALHO, J.M., MURRY, M.H., WURDEMAN, K. et KIM, J. (1993). Estimating heterogeneity in the probabilities of enumeration for *Association*. 88, 1130-1136.

BREIMAN, L., FRIEDMAN, J.H., OLSEN, R.A. et STONE, C.J. (1984). *Classification and Regression Trees*. Wadsworth, California.

BROOKS, S.P., CATCHPOLE, E.A. et MORGAN, B.J.T. (2000). Bayesian animal survival estimation. *Statistical Science*. 15, 357-276.

BROOKS, S.P., et GELMAN, A. (1998). Alternative methods for monitoring convergence of iterative simulation. *Journal of Computational and Graphical Statistics*. 7, 434-455.

CHRISTIANSEN, C.T., et MORRIS, C. (1997). Hierarchical Poisson regression models. *Journal of the American Statistical Association*. 92, 618-632.

DI PIETRO, E. (1998). Anagrafi comunali: funzione statistica e livello di informatizzazione. *Atti Della Quarta Conferenza Nazionale di Statistica*. Tomo I - Sessioni Plenarie, Workshop: Il progetto anagrafi. Roma. 11-13.

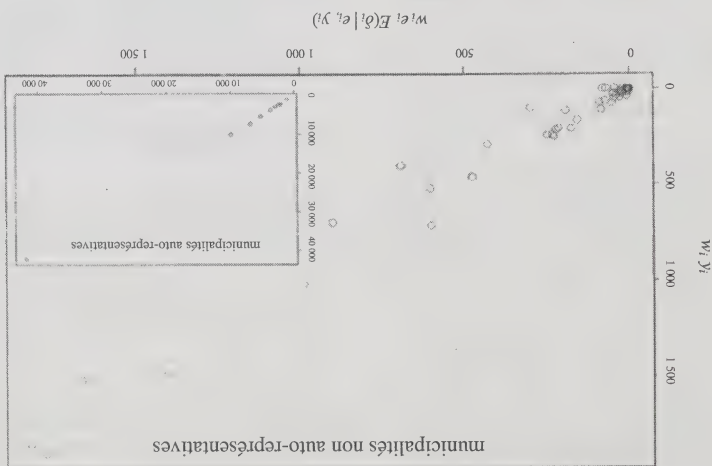


Figure 4. Estimations composites en fonction des estimations directes du nombre de ménages non dénombrés dans chaque municipalité.

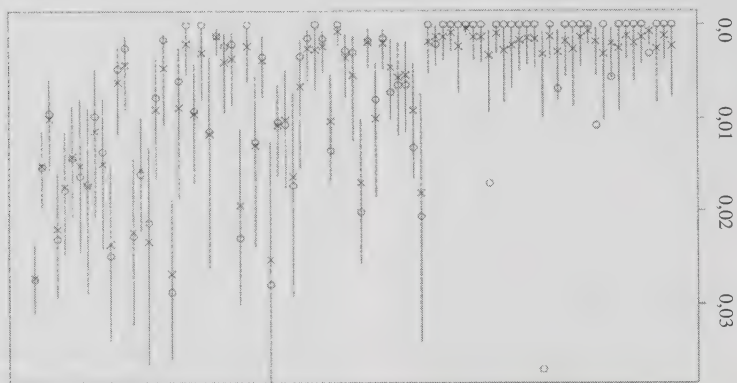


Figure 5. Estimations composites (x) et leurs intervalles de crédibilité à 95 %; (y) estimations directes. Les municipalités sont triées par taille de population.

REMERCIEMENTS

Les données de l'EPR et les archives contenant les données sur les municipalités ont pu être utilisées grâce à une entente spéciale entre l'ISTAT et le Département de

statistique de l'Université de Bologne.

Nous tenons à remercier Francesca Bruno et Loredana Di Consiglio de leur contribution inestimable à la préparation des ensembles de données de base, et Meri Raggi de son soutien permanent et de sa discussion des

sujets de la présente étude.

Nous remercions le rédacteur, un rédacteur adjoint et deux examinateurs anonymes de leurs commentaires et suggestions qui nous ont aidé à réviser et à améliorer le

manuscrit.

Les travaux ont été financés en partie par la subvention du projet de recherche « Quality of total and partial surveys » (1999-2000) de l'Université de Bologne (60 %).

Nous remercions Angela Ferruzzi, Marco Fortini, Aldo Orasi et Fernanda Panizon, de l'équipe d'ISTAT travaillant sur le Recensement et l'EPR de 2001, ainsi que Martella Dimitti et Ersilia Di Pietro, du groupe d'ISTAT s'occupant des enquêtes sur le rendement statistique des municipalités, de leurs suggestions utiles et de leur soutien continu.

Tableau 1
AIC des modèles estimés comparativement au modèle de référence M_0

Variables dans les modèles									
Effets de groupe + variables de qualité et démographiques		Effets de variables démographiques		Effets de variables de qualité		Effets de groupe + variables de qualité et démographiques		Effets de groupe + variables de qualité et démographiques	
Région	-4,22	15,34	2,08	6,13	17,87	-0,39	18,52	23,32	20,09
Classes de pop. mun.								8,45	17,83
Effets de Région* classes pop. mun.								4,91	13,74
Région + classes pop. mun.								23,48	26,15
Arbre 2 (variables de qualité)	11,81			8,34				35,53	41,45
Arbre 1 (variables démographiques)	35,14			35,37				32,28	
Arbre 3 (variables de qualité + démographiques)	38,89			35,76				41,12	

Nous avons de nouveau estimé l'ensemble de modèles après élimination de la municipalité la plus grande, qui constitue un cas influent éventuel. De nouveau, le modèle fondé sur l'arbre 3 avec variables démographiques et de qualité en tant que variables explicatives a été sélectionné (l'aide du critère (11). Ce modèle est caractérisé par un bon ajustement (la valeur p bayésienne associée à la mesure de divergence (12) est égale à 0,51). En outre, les estimations composées varient peu comparativement à celles obtenues au moyen de l'échantillon entier.

Afin de vérifier l'ajustement du modèle, à la figure 4, nous avons tracé les estimations composées en fonction des estimations directes du nombre de ménages non dénombrés dans chaque municipalité (les valeurs pour les 10 plus grandes municipalités sont présentées pour une échelle différente). Les estimations composées sont $w_i e_i E(\delta_i | y_i, e_i)$, tandis que les estimations directes sont $w_i y_i$, w_i étant le facteur d'expansion dû à l'échantillonnage des SD dans chaque municipalité. Les estimations composées sont les espérances a posteriori des paramètres de premier niveau et, conditionnellement aux hyperparamètres, sont des estimations composées fondées sur le modèle représenté par les λ_i lorsqu'on dispose de données d'échantillonnage valables. D'après (5), nous savons que cette méthode de pondération est régie par les facteurs municipaux de rétrécissement B_i . Ceux-ci pondèrent les estimations directes y_i/e_i proportionnellement à e_i/λ_i , c'est-à-dire le nombre de ménages non dénombrés dans l'échantillon municipal prévu par le modèle.

Pour les municipalités comptant jusqu'à 10 000 résidents (cette valeur est assez proche de la valeur de partition de 13 200 de l'arbre 3), dans presque tous les cas, nous avons des valeurs de B_i très proches de 1; autrement dit, pour les petites municipalités, la composante basée sur le modèle joue un rôle prédominant dans la détermination de l'estimation composite. À la figure 5, les estimations composées

Les résultats de la présente étude, où l'on envisage pour les petites municipalités sélectionnées. Les résultats de la présente étude, où l'on envisage pour les petites municipalités sélectionnées. Les résultats de la présente étude, où l'on envisage pour les petites municipalités sélectionnées.

en moyenne, à titre de « rodage » prudent, ce qui donne environ 20 000 tirages à partir de la loi a posteriori de chaque modèle.

5. COMPARAISON DES MODÈLES ET DISCUSSION DES RÉSULTATS EMPIRIQUES

Nous avons estimé une gamme de modèles pour diverses définitions des matrices des variables explicatives X et Z . En ce qui concerne la matrice de plan d'expérience Z , nous considérons sept cas distincts, dans lesquels les municipalités sont groupées en fonction des critères de stratification classiques (région géographique et taille de la population) ou des résultats des partitions techniques décrites à la section 3. Il s'agit des groupements fondés sur :

- a) la région géographique (Nord, Centre, Sud et Îles), b) les classes de taille de population uniquement, c) les classes de taille de population selon la région géographique, d) les classes de taille de population et les régions géographiques, e) l'arbre 1 (fondé sur les variables démographiques), f) l'arbre 2 (fondé sur les variables de qualité), g) l'arbre 3 (fondé sur les variables démographiques et de qualité). On peut proposer deux types de variables pour la matrice X , à savoir les variables de qualité de la section 2.2 et les variables démographiques de la section 2.3. Par conséquent, la matrice X peut avoir trois compositions distinctes :
 - I) variables de qualité uniquement, II) variables démographiques uniquement et III) variables de qualité et diverses définitions de X et Z , nous avons estimé vingt-huit modèles distincts. Cette méthode nous permet d'introduire différents blocs de variables, au lieu de procéder à la sélection de variables.

La quantité habituellement utilisée pour comparer les modèles dans le cadre bayésien est le facteur de Bayes (BF). Une approximation en grand échantillon de $-2 \ln(BF)$ est donnée par

$$\Delta BIC = -2 \ln \left[\frac{\sup_{M_0} f(y | \theta_0)}{\sup_{M_k} f(y | \theta_k)} \right] - (p_k - p_0) \ln n \quad (11)$$

(voir Schwarz 1978) qui, de surcroît, ne renvoie à aucune des hypothèses a priori. Nous notons que, dans (11), les $M_k (k = 1, \dots, K)$ indiquent l'ensemble de modèles concurrents et que θ_k est le paramètre dimensionnel p_k indiquant la vraisemblance associée à chaque modèle. Le modèle nul auquel sont comparés tous les autres, qui est celui comportant la seule coordonnée à l'origine, est dénoté par M_0 . Les valeurs positives et grandes de (11) appuient le modèle M_k .

Dans (11), la pénalisation de complexité dépend de la taille du sous-ensemble de paramètres de troisième niveau; autrement dit, tous les modèles sont comparés comme s'ils n'étaient pas hiérarchiques. Puisqu'ils ont une même structure hiérarchique, cette modification opérationnelle du

critère d'information bayésien type BIC ne modifie pas les résultats de la comparaison des modèles résumés au tableau 1.

Nous notons que les modèles où les effets de groupe sont fondés sur la région géographique donnent de forts mauvais résultats (ligne 1) et qu'il en est de même si l'on combine la région géographique et la taille de la population des municipalités (lignes 3 et 4). Ces résultats sont assez étonnants, puisqu'on se fonde sur les régions géographiques pour concevoir la stratification de l'échantillon de l'EPR et que l'efficacité des administrations, ainsi que d'autres indicateurs socio-économiques sont censés être regroupés en fonction des grandes subdivisions géographiques de l'Italie (Nord, Centre, Sud). Ce résultat peut être attribué au rôle prédominant que l'organisation particulière de chaque municipalité joue dans la détermination de l'efficacité des opérations du recensement sur le territoire de la municipalité.

Les modèles à effets de groupe fondés sur un arbre (lignes 5 à 7) donnent de nettement meilleurs résultats que ceux à effets de groupe fondés sur les critères de stratification habituels d'ISTAT (lignes 1 à 4). Seuls font exception les modèles basés sur l'arbre 2 (ligne 5), qui donnent des résultats assez médiocres lorsqu'on n'inclut pas la taille de la population et d'autres variables démographiques. En fait, la population municipale peut être considérée comme étant une approximation de la complexité organisationnelle de la municipalité. Il semble que les variables de qualité soient des discriminants puissants du niveau de sous-décomposément parmi les municipalités dont les caractéristiques démographiques sont semblables, mais qu'ils aient peu de pertinence si l'on ne tient pas compte de l'effet d'un degré de complexité organisationnelle différent par introduction d'une variable de taille de population. Nous soulignons que l'ajout d'une matrice de plan d'expérience Z fondée sur des groupes de municipalités établis d'après des arbres de régression de Poisson nous permet de modéliser les relations non linéaires entre le sous-décomposément et les variables explicatives.

En fait, les modèles fondés sur l'arbre 3 sont ceux dont les propriétés sont les meilleures. Suivent plusieurs commentaires sur le modèle dont le ΔBIC est maximal. Ce modèle comprend des variables démographiques et des variables de qualité comme variables explicatives. L'adéquation du modèle choisi est évaluée au moyen de vérifications prédictives a posteriori. Plus précisément, nous adoptons la mesure d'usage générale de la divergence par rapport à l'ajustement valide du modèle proposée par Brooks, Catchpole et Morgan (2000) en tant qu'outil approprié pour les occurrences rares, comme les sous-

dénombréments au recensement :

$$D(y; \theta) = \sum_{i=1}^I \left(\sqrt{y_i} - \sqrt{\text{Exp}_i} \right)^2, \quad (12)$$

où $\text{Exp}_i = e_i^T E(\delta_j | y_i, e_i)$. La probabilité de 0,46 associée à l'aire de la queue témoigne d'un bon ajustement pour le modèle choisi.

$$(6) \quad \beta_j \stackrel{\text{ind}}{\sim} N(0, 100), \quad j = 1, \dots, p$$

$$(7) \quad \xi_k \stackrel{\text{ind}}{\sim} N\left(k u_k, \frac{1}{n_k}\right), \quad k = 1, \dots, q$$

où \bar{n}_k est le nombre moyen de ménages échantillonnés dans les n_k municipalités du même groupe. Les lois a priori (7), associées aux effets de groupe, sont par conséquent centrées autour de la taille des groupes. Elles sont construites de façon à être faiblement informatives en vue d'améliorer les propriétés de stabilité et de convergence du modèle. Les lois a priori des coefficients de régression (6) associées aux autres variables explicatives sont centrées sur 0. Pour le paramètre de surdispersion ζ , nous choisissons la loi a priori

$$(8) \quad \zeta \sim 1\,000 \cdot \text{Gamma}(0,001, 1)$$

en suivant la proposition de Christensen et Morris (1997).

Notons que les deux premiers moments a priori de (8) sont $E(\zeta) = 1$ et $V(\zeta) = 1\,000$; donc, la loi a priori est très diffuse et caractérisée par une forte asymétrie positive.

Au quatrième niveau de la hiérarchie, nous spécifions les lois a priori suivantes :

$$(9) \quad k \sim N(0, 100)$$

$$(10) \quad \tau \sim \text{Gamma}(0,001, 0,001).$$

qui sont toutes deux conçues pour avoir un effet très faible sur les inférences a posteriori.

Nous calculons les lois a posteriori de $(\delta_i | y_i, e_i)$ à l'aide d'algorithmes d'échantillonnage de Monte Carlo par chaînes de Markov (MCMC). Pour ces calculs, nous utilisons le logiciel BUGS (Spiegelhalter, Thomas, Best et Gilks 1995), qui est fondé sur l'échantillonnage de Gibbs.

Puisque la résolution des modèles comportant des lois discrètes demande des calculs compliqués, nous spécifions les lois a priori (6) à (10) en choisissant des formes fonctionnelles simples bien connues, comme la loi normale et la loi Gamma, qui facilitent les calculs. L'examen de la sensibilité des moyennes a posteriori dans (6) à (10) n'a

révélé aucune variation importante de ces moyennes. Donc, nous pouvons considérer les lois a priori comme étant non informatives. Pour évaluer la convergence, nous considérons la méthode à chaînes multiples proposée par Gelman et Rubin (1992), et nous exécutons trois chaînes différentes avec point de départ bien distinct pour chaque modèle.

Nous considérons l'inspection visuelle du cheminement des chaînes et la statistique modifiée de Gelman et Rubin (Brooks et Gelman 1998) comme des outils élémentaires d'évaluation de la convergence. Nous avons exécuté 10 000 itérations pour chaque chaîne, en éliminant 3 000

Au lieu de la paramétrisation susmentionnée, nous adoptons celle de la loi Gamma au deuxième niveau de la hiérarchie, conformément à la proposition de Christensen et Morris (1997). Si nous supposons

$$(4) \quad \delta_i | \lambda_i, \zeta \sim \text{Gamma}(\zeta, \zeta/\lambda_i)$$

avec les moments $E(\delta_i | \lambda_i, \zeta) = \lambda_i$ et $V(\delta_i | \lambda_i, \zeta) = \lambda_i^2/\zeta$, nous obtenons

$$y_i | e_i, \lambda_i, \zeta \sim \text{NegBin}\left(\frac{\zeta/\lambda_i}{\zeta/\lambda_i + e_i}, \zeta, \frac{\zeta/\lambda_i}{\zeta/\lambda_i + e_i}\right),$$

où $V(y_i | e_i, \lambda_i, \zeta) - E(y_i | e_i, \lambda_i, \zeta) = e_i^2 \lambda_i^2/\zeta$. À mesure que ζ tend vers l'infini, la variance de la loi binomiale négative converge vers celle de la loi de Poisson (la variance de la loi Gamma en (4) tend vers 0), tandis que les valeurs faibles de ζ indiquent une forte surdispersion.

Partant de (4), nous voyons immédiatement que :

$$E(\delta_i e_i | e_i, \lambda_i, \zeta) = \lambda_i e_i;$$

donc, l'hypothèse de dépendance (3) se réécrit en fonction de $\lambda_i e_i$ sous la forme :

$$\ln(\lambda_i e_i) = X_i' \beta + Z_i' \zeta.$$

La loi a priori (4) est conjuguée à la vraisemblance définie par (2). Conséquemment, nous obtenons

$$\delta_i | y_i, e_i, \lambda_i, \zeta \sim \text{Gamma}(y_i + \zeta, e_i + \zeta/\lambda_i)$$

dont il découle que

$$(5) \quad E(\delta_i | y_i, e_i, \lambda_i, \zeta) = (1 - B_i) r_i + B_i \lambda_i$$

$$\text{où } r_i = y_i/e_i \text{ et } B_i = \zeta/(\zeta + e_i \lambda_i).$$

Nous pouvons considérer chaque moyenne a posteriori (5) comme un estimateur sur petit domaine composite où les composantes directes et synthétiques sont toutes deux pondérées conformément à l'information fournie par l'échantillon.

D'après (5), nous notons que la moyenne a posteriori de la distribution des paramètres de taux δ_i est une combinaison linéaire du taux de sous-dénombrement observé r_i et de la moyenne a priori λ_i . Autrement dit, le modèle présente une linéarité a posteriori. Dans (5), les deux termes sont pondérés d'après B_i , dont la valeur varie entre 0 et 1. Plus la valeur de B_i est grande, plus le poids de la moyenne a priori λ_i (estimateurs synthétiques) est grande et les estimations produites par le modèle prennent de l'importance comparativement aux taux observés. Nous constatons que chaque B_i est inversement proportionnel au terme $e_i \lambda_i$, qui exprime la quantité d'information que fournit l'échantillon de chaque domaine.

Pour achever la spécification bayésienne du modèle, nous attribuons une loi aux paramètres de troisième niveau ζ, β, ζ . D'après un critère approximatif non informatif, nous introduisons des lois a priori correctes, mais plates. Plus précisément, nous supposons que :

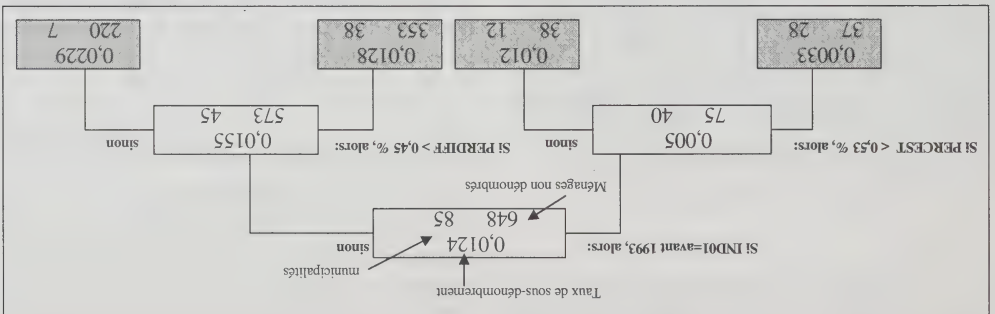


Figure 2. Arbre 2 fondé sur les variables de qualité des statistiques des municipalités.

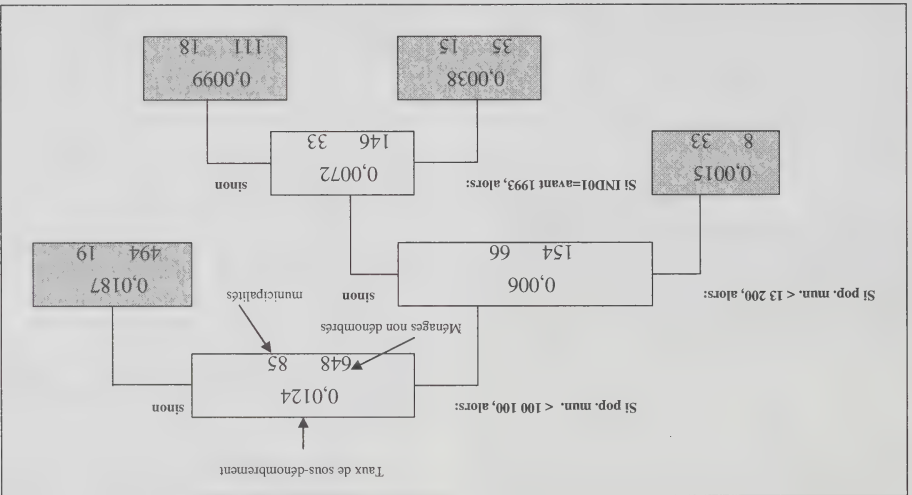


Figure 3. Arbre 3 fondé sur des variables démographiques et de qualité.

4. MODELES POISSON-GAMMA HIERARCHIQUES

Nous représentons le nombre de ménages non dénombrés observé dans chaque échantillon municipal par $y_i (i = 1, \dots, 85)$. À titre de première approximation, nous pouvons modéliser ces dénombrements par une loi de Poisson :

$$y_i | \delta_i, e_i \sim \text{Pois}(\delta_i e_i) \quad (2)$$

où δ_i représente le taux de sous-dénombrement qu'il faut estimer et e_i est donné par le nombre de ménages dans les SD échantillonnés dans la municipalité. Nous exprimons la dépendance à un ensemble de variables explicatives par un lien canonique logarithmique :

$$\ln(\delta_i e_i) = X_i^T \beta + Z_i^T \xi \quad (3)$$

où Z_i est la i^{e} ligne d'une matrice nominale de plan d'expérience introduite pour modéliser les effets de groupe. Chaque X_i^T est un vecteur de dimension p de variables explicatives associées à la i^{e} municipalité, et β et ξ sont les paramètres de régression.

Comparativement au nombre de ménages observés, les cas de non-dénombrement sont assez rares. Par conséquent, les données peuvent présenter une surdispersion importante. Le problème de la surdispersion peut être résolu par modélisation hiérarchique des paramètres δ_i dans (2). Si les paramètres δ_i suivent une loi Gamma(α, ν), on obtient marginalement la loi binomiale négative pour y_i par intégration des paramètres δ_i , i.e. $y_i | \alpha, \nu, e_i \sim \text{NegBin}(\alpha, \nu / (\nu + e_i))$ avec les moments :

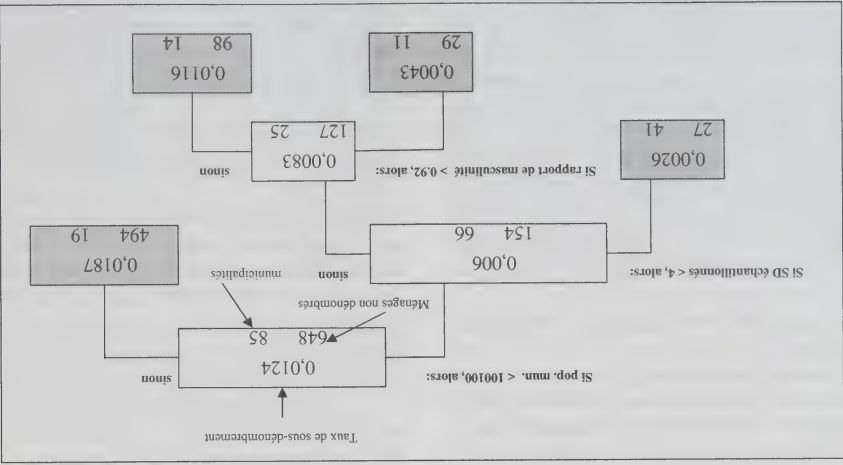
$$E(y_i | \alpha, e_i, \nu) = \frac{\alpha e_i}{\alpha e_i + \nu}, \quad V(y_i | \alpha, e_i, \nu) = \frac{\alpha e_i (\nu + e_i)}{\nu^2} \quad (\text{voir Lawless 1987}).$$

partition séparé les municipalités de moins de 100 100 habitants de celles de plus de 100 100 habitants. Cette valeur de partition coïncide presque avec le seuil de démarcation de 100 000 utilisé pour la stratification des municipalités lors de l'EPR de 1991. La deuxième partition isole un sous-échantillon de petites municipalités pour lesquelles moins de quatre SD ont été échantillonnées pour l'EPR. Une partition supplémentaire est faite d'après le rapport de masculinité.

L'arbre 2 (figure 2) est fondé exclusivement sur des variables concernant la qualité des statistiques des municipalités. La première partition s'appuie sur le temps mis pour corriger les registres de population (IND01) : les municipalités qui procèdent le plus rapidement à cette tâche affichent les taux de sous-dénombrement les plus faibles. Les partitions de niveau inférieur mettent en relief le problème des personnes résidant provisoirement à l'étranger (PERCEST), problème qui, dans les régions caractérisées par une émigration massive, peut donner lieu à un sous-dénombrement grave de la population municipale et à des erreurs de tenue à jour des registres de population (PERDIFE). Dans cet arbre, une moitié de l'échantillon est classée dans un seul noeud qui contient vraisemblablement l'hétérogénéité résiduelle.

L'arbre 3 (figure 3) est fondé sur des variables à la fois démographiques et de qualité. La première partition est basée sur la population municipale, exactement de la même façon que pour l'arbre 1. Subséquentement, le sous-ensemble de municipalités comptant moins de 100 100 habitants est partitionné en un groupe de municipalités de petite taille et un groupe de municipalités de taille moyenne au seuil de 13 200 habitants. La variable de qualité incluse dans cet arbre est le temps mis pour corriger les registres de population (IND01).

Figure 1. Arbre 1 fondé sur les variables démographiques.



$$D_T^j(\alpha) = D_T^j + \alpha \text{size}(T^j) \quad (1)$$

Pour un α spécifié, on peut trouver l'arbre $T(\alpha)$ qui minimise (1). On peut montrer (Breiman et coll. 1984) qu'il existe une famille emboîtée de sous-arbres $\{T_0, T_1, \dots, T_{T^{\text{root}}}, \dots, T_{T^{\text{root}}}\}$ de T_0 telle que chaque arbre est optimal pour une fourchette de valeurs de α .

Le problème se réduit maintenant à choisir l'un de ces sous-arbres. La sélection se fait de façon à minimiser l'erreur de prédiction définie comme étant la contribution d'une nouvelle observation à la somme des carrés des écarts. Pour estimer l'erreur de prédiction, l'existence d'un échantillon indépendant serait, théoriquement, la meilleure option, mais puisqu'il est préférable d'utiliser toutes les données afin d'"informer" l'arbre le mieux possible, on choisit l'arbre T_0 dont l'erreur de prédiction estimée est minimale. Ici, nous utilisons une règle d'élagage plus stricte qui consiste à sélectionner le plus petit arbre ayant une erreur estimée de prédiction n'excédant pas la somme de l'erreur estimée de prédiction de T_0 et de son erreur-type. L'erreur estimée de prédiction de T_0 et de son erreur-type. Nous adoptons cette règle d'élagage, appelée « règle d'une erreur-type » (Breiman et coll. 1984), pour éviter de surajuster le modèle.

Puisque la contre-vérification des arbres de régression de Poisson peut donner, pour certains noeuds, une valeur infinie de la somme des carrés des écarts, nous utilisons des estimateurs bayésien de réécissement des taux réels fondés sur un simple modèle Poisson-Gamma, comme l'ont proposé Therneau et Atkinson (1997).

Nous construisons trois arbres distincts en partant de l'arbre 1 (présenté à la figure 1) s'appuie sur les variables démographiques uniquement. La première

d'échantillonnage ont été sélectionnées avec probabilités égales par échantillonnage systématique. L'échantillon final de l'EPR contenait 85 municipalités et 638 SD (sur un total national de 8 095 municipalités et 64 000 SD) avec une estimation fondée sur un plan d'échantillonnage national de 1,24 % (Abbate, Masselli et Signore 1999).

Le questionnaire de l'EPR, qui ne contient que quelques questions simples, a été rempli au cours d'une interview sur place. Les caractéristiques des ménages échantillonnés sont limitées au nombre et au sexe des membres du ménage. D'autres questions de l'EPR ont été conçues pour faciliter le couplage des enregistrements aux données du recensement, donc, pour réduire le nombre de cas de dénombrement à mauvais endroit et d'autres erreurs non dues à l'échantillonnage lors de l'évaluation du sous-dénombrement (voir Fortini 1994, pour plus de précisions).

2.2 Enquêtes sur la qualité des statistiques des municipalités

ISTAT a créé un ensemble de données sur la qualité des statistiques des municipalités italiennes (voir Di Pietro 1998, 1999). Cet ensemble intègre différentes sources, dont l'information provenant des enregistrements sur le recensement lors du Recensement de 1991, des registres de population municipaux et de données du ministère de l'Intérieur. Cet ensemble de données contient aussi les résultats des trois enquêtes administratives réalisées durant les années 1990 en vue d'évaluer le rendement des municipalités en ce qui concerne leurs engagements à l'égard d'ISTAT. La première enquête porte sur l'information des bureaux municipaux de la statistique. La deuxième, connue sous l'acronyme POSAS, est une enquête post-censitaire (EP) fondée sur les registres de la population de résidents, classés selon l'année de naissance, l'âge et l'état civil. La troisième, connue sous l'acronyme ISCAN, vise à déterminer la mesure dans laquelle les enregistrements figurant sur les listes des registres municipaux de population sont appropriés. Ces enquêtes fournissent des données sur toutes les municipalités italiennes.

À partir de cet ensemble de données, nous avons sélectionné un sous-ensemble de variables reliées à l'activité municipale au moment du Recensement de 1991, à savoir :

- a) le pourcentage de champs non codés du questionnaire du recensement à l'intention des ménages qui auraient dû être remplis, après l'interview des ménages, par les bureaux municipaux de la statistique (FERCOD);
- b) le ratio de la population provisoirement à l'étranger à la population présente au moment du Recensement de 1991 (PERCEST);
- c) le ratio de la différence entre les dénombrements du Recensement de 1991 et ceux des registres de population aux dénombrements du Recensement de 1991 (PERDIFF);

- d) le temps nécessaire pour mettre à jour les registres de population municipaux d'après les résultats du Recensement de 1991 (IND01);
 - e) le retard de la mise à jour des noms de rue (IND11).
- ## 2.3 Variables démographiques
- Nous considérons aussi l'ensemble de ratios démographiques établis d'après les résultats du Recensement de 1991. En particulier, nous utilisons les pourcentages de ménages « ne comptant qu'un seul membre » et « comptant plus d'une famille », ainsi que les ratios de masculinité (ratios hommes-femmes) dans la municipalité. La population municipale de résidents – correspondant aux chiffres non corrigés du Recensement de 1991 – est aussi une variable fort importante. Le nombre de SD échantillonnés dans chaque municipalité pour l'EPR est un autre indicateur de l'importance de la municipalité.

3. ARBRES DE RÉGRESSION DE POISSON

Les sources de données disponibles nous fournissent un grand nombre de variables auxiliaires, dont beaucoup sont nominales ou polychotomiques. Avant d'ajuster les modèles hiérarchiques, nous regroupons les municipalités dont le taux de sous-dénombrement des ménages est homogène à l'aide d'arbres de régression binaires de Poisson. Les groupes établis d'après ces arbres sont inclus à titre de facteurs dans les modèles décrits à la section suivante. Notre objectif principal est de vérifier l'efficacité des méthodes habituelles de stratification, de les améliorer *ex post* à l'aide de modèles hiérarchiques contenant des covariables appropriées et de comparer les résultats ainsi obtenus à ceux de méthodes semblables fondées sur des groupements optimaux.

Les modèles de régression conditionnelle sont fondés sur le lien logarithmique canonique. Le critère de partition est la statistique habituelle de somme des carrés des écarts, ou *deviance* en anglais (Therneau et Atkinson 1997) :

$$Deviance_{parent} - (Deviance_{enfant, gauche} + Deviance_{enfant, droite})$$

L'idée fondamentale de la construction d'un arbre consiste à partir d'un grand arbre T_0 construit en appliquant une règle d'arrêt naïve et faible (comme le nombre minimal d'observations dans les nœuds finaux de l'arbre), puis à sélectionner par élagage l'arbre de taille appropriée parmi les sous-arbres de T_0 . La méthode établie pour élaguer les arbres est celle du coût-complexité, introduite pour la première fois par Breiman, Friedman, Olshen et Stone (1984). Soit D_T la somme des carrés des écarts d'un sous-arbre T de T_0 , le nombre de nœuds terminaux de T et $\alpha > 0$ un paramètre de coût-complexité :

ménages, puisques nombre de caractéristiques déterminant la proposition individuelle à être dénombré lors du recensement s'agissent lorsqu'on analyse des données agrégées. Dans le cas de l'Italie, une analyse complète, fondée sur des questionnaires de l'EPR de 1991 ne contenait qu'un très petit nombre de questions s'adressant à des individus. Par conséquent, l'EPR de 1991 fournit fort peu de données auxiliaires sur les SD, de sorte qu'on ne peut proposer des modèles fondés sur le sous-dénombrement au niveau des SD. Notre analyse se fonde sur la combinaison de différentes sources de données. Les données auxiliaires proviennent de l'EPR de 1991 susmentionnée, de deux études sur la qualité des statistiques des municipalités réalisées par ISTAT au cours des années 1990 (DI Pietro 1998, 1999) et sur des indicateurs démographiques et sociaux tirés des résultats officiels du Recensement de 1991.

Le problème est de savoir comment utiliser efficacement l'information provenant des diverses sources de données. Nous disposons en fait d'un grand nombre de variables, dont la plupart sont nominales ou polychotomiques. Au lieu d'utiliser un algorithme de sélection de variables, nous avons décidé de former des groupes homogènes de municipalités, puis de les introduire dans le modèle à l'aide d'une matrice de plan d'expérience pour les effets aléatoires. Nous utilisons, pour construire ces groupes, des arbres de régression de Poisson (Therneau et Atkinson 1997). Cette utilisation hiérarchique de l'information constitue une base naturelle pour la formation de strates de municipalités géographiquement non contiguës.

Durant l'EPR, peu de SD sont de nouveau dénombrés à l'intérieur de chaque municipalité échantillonnée; le taux moyen d'échantillonnage des SD est de 0,001. Il s'agit d'un contexte typique d'estimation sur petits domaines où les estimations directes du taux municipal de sous-dénombrement ne sont pas fiables et doivent être remplacées par des estimations synthétiques ou composites fondées sur un modèle approprié. Le phénomène du sous-dénombrement est rare. Nos données sont constituées de dénombrements et peuvent présenter une surdispersion importante par rapport à l'hypothèse d'une loi de Poisson. Nous proposons l'utilisation de modèles de régression hiérarchiques de Poisson pour tenir compte de la surdispersion.

Les modèles hiérarchiques adoptés ici permettent de traiter explicitement la surdispersion à cause de l'hétérogénéité municipale. Une autre source de variabilité extra-poissonienne est due à l'hétérogénéité à l'intérieur des municipalités, à cause de l'existence de grappes de personnes non dénombrées dans les SD ou des grappes dues aux personnes non dénombrées dans une même famille. Cette forme de surdispersion n'est pas traitée explicitement dans les modèles. Nous adoptons une approche entièrement bayésienne pour la spécification et l'estimation, et nous résolvons les modèles par les méthodes de simulation de Monte Carlo par chaînes de Markov. Dans ce cadre hiérarchique, nous

traçons la surdispersion en imposant une loi Gamma pour les taux du premier niveau de la loi de Poisson, donc en obtenant marginalement une loi binomiale négative. En outre, conditionnellement aux hyperparamètres, le modèle proposé présente une linéarité a posteriori et les moyennes a posteriori correspondantes des taux de sous-dénombrement municipaux sont des estimations linéaires complètes. Donc, le degré de lissage dépend de la quantité d'information fournie par chaque échantillon municipal dans l'EPR.

Nos résultats montrent qu'on pourrait améliorer la stratification des municipalités sur laquelle repose le plan de sondage de l'EPR de 1991 (fondé sur la région géographique et la taille de population), puisque le taux de sous-dénombrement est en grande partie indépendant de la région géographique. Par contre, les variables décrivant l'efficacité statistique des administrations locales aident à faire la distinction entre les divers degrés de sous-dénombrement parmi les municipalités de mêmes taille et structure démographique. Si l'on ne modifie pas le plan de sondage de l'EPR, nos résultats pourraient fournir des indications utiles lors de l'analyse des données.

Le présent article est présenté comme suit. À la section 2, nous décrivons les caractéristiques générales de l'EPR et des autres sources de données utilisées. À la section 3, nous examinons les arbres de régression de Poisson employés pour créer des groupes homogènes de municipalités. À la section 4, nous introduisons les modèles de régression de Poisson hiérarchiques, tandis qu'à la section 5, nous discutons des résultats empiriques et comparons les modèles.

2.1 Enquête postcensitaire (EP) de l'Italie

2. DONNÉES DE L'EPR ET INFORMATION AUXILIAIRE

Le Recensement de la population de l'Italie de 1991 a eu lieu le 20 octobre. L'Enquête postcensitaire (EP), fondée sur un plan d'échantillonnage stratifié à deux degrés, a été réalisée quelques semaines plus tard. Les municipalités représentent les unités primaires d'échantillonnage et les secteurs de dénombrement, les unités secondaires. Un SD est le plus petit domaine en lequel le territoire municipal est divisé aux fins des opérations du recensement; chaque SD est affecté entièrement à un même intervieweur. Les unités primaires d'échantillonnage ont été stratifiées en fonction de la région géographique (Nord-Ouest, Nord-Est, Centre, Sud, îles) et de la taille de la population (sept classes pour les municipalités comptant moins de 350 000 habitants), ce qui a produit 35 strates. Dans chaque strate, les municipalités échantillonnées ont été sélectionnées sans remise et avec probabilité proportionnelle à la taille de la population. Les 10 municipalités comptant plus de 350 000 habitants ont été incluses dans l'échantillon en tant qu'unités auto-représentatives. Les unités secondaires

Un modèle hiérarchique pour l'analyse du sous-dénombrement local du recensement en Italie

D. COCCHI, E. FABRIZI et C. TRIVISANO¹

RÉSUMÉ

La comparaison des résultats des recensements et des enquêtes postcensitaires (EP) montre que les chiffres de recensement sont inexacts. En Italie, les administrations municipales jouent un rôle essentiel dans les opérations sur le terrain du recensement et de l'EP. Dans le présent article, nous analysons l'effet des municipalités sur le taux de sous-dénombrement au recensement en Italie par modélisation des données provenant de l'EP et d'autres sources à l'aide d'arbres de régression de Poisson et de modèles de Poisson hiérarchiques. Les arbres de régression de Poisson permettent de former des groupes homogènes de municipalités. Les modèles de Poisson hiérarchiques peuvent être considérés comme des outils pour l'estimation pour des petits domaines.

MOTS CLÉS : Sous-dénombrement au recensement; enquête postcensitaire (EP); modélisation bayésienne hiérarchique; modèles de régression Gamma-Poisson; arbres de régression de Poisson.

1. INTRODUCTION

Le recensement de la population de l'Italie, qui a lieu tous les dix ans, représente la tâche la plus importante de l'Institut national de statistique de l'Italie (ISTAT). Les travaux qui ont abouti au présent article ont été réalisés juste avant le Recensement de l'Italie de 2001 et l'EPR subséquent. On a tenu compte des résultats obtenus pour la réalisation de l'EPR de 2001). Pour procéder au recensement, ISTAT s'appuie sur les administrations municipales qui sont chargées de toutes les opérations sur le terrain (formation des intervieweurs, planification des interviews, collecte et traitement de base des données). Durant les opérations de recensement, les municipalités travaillent indépendamment les unes des autres sous la supervision d'ISTAT. Par conséquent, l'exactitude des résultats varie considérablement d'une municipalité à l'autre, même si elles sont contiguës. En Italie, le territoire d'un bourg municipal est subdivisé en plusieurs secteurs de dénombrement (SD), qui sont affectés à un même intervieweur durant les opérations du recensement. Les SD diffèrent par leur forme, leur structure et la difficulté de dénombrement, ainsi que par l'intervieweur. Aussi est-il probable que le taux de sous-dénombrement varie fortement d'un SD à l'autre dans une même municipalité.

Après le Recensement de la population de 1991, ISTAT a réalisé une Enquête postcensitaire (EP) pour évaluer le phénomène du sous-dénombrement. Il est bien connu que les chiffres de population du recensement sont généralement incorrects parce que des personnes ne sont pas dénombrées, sont dénombrées plusieurs fois ou sont dénombrées au mauvais endroit. Les personnes non dénombrées sont la source la plus importante d'inexactitude et, habituellement, produisent un sous-dénombrement net qui peut varier selon la région géographique ou selon le

groupe social, et influencer la détermination de la taille relative des sous-populations (Abbate, Masselli et Signore 1993). Les opérations sur le terrain de l'EPR de 1991 ont été réalisées par les municipalités échantillonnées directement dites. Les données recueillies ont été analysées par Abbate, Masselli et Signore (1993), qui ont estimé le taux national global de sous-dénombrement au moyen d'un modèle de Lincoln-Petersen (voir Wolter 1986) en utilisant des strates à posteriori de municipalités fondées sur de grandes régions géographiques (Nord, Centre, Sud). À partir des mêmes données, Fortini (1994) a estimé le sous-dénombrement national global au moyen de modèles de classes latentes.

Au lieu d'estimer le taux de sous-dénombrement pour l'ensemble du pays ou pour des domaines plus petits, nous proposons des modèles conçus pour expliquer la variation du taux de sous-dénombrement au niveau municipal. La découverte de facteurs expliquant la grandeur du sous-dénombrement net pourrait servir de base à la création de groupes homogènes de municipalités qui permettraient de mieux planifier la stratification lors de futures enquêtes postérieures au recensement. En outre, le dépistage des faibles dans la structure organisationnelle qui influent de façon significative sur le sous-dénombrement pourrait fournir des indications quant aux mesures à prendre pour réduire l'importance de ce dernier.

On trouve dans la littérature des études fondées sur des données d'EPR désaggrégées. Alho, Mulry, Wurdeman et Kim (1993) considèrent un modèle de régression logistique pour la probabilité individuelle (ménage) d'être dénombré. En nous inspirant de Moura et Holt (1999), nous pourrions étendre le modèle afin d'y inclure les effets de municipalité ou d'autres groupes. Nous sommes parfaitement conscients que notre décision de modéliser des données municipales n'équivaut pas à analyser les enregistrements au niveau des

- MULRY, M.H., et SPENCER, B.D. (1988). L'erreur totale dans l'estimation de système dual : Recensement du Central Los Angeles County de 1986. *Techniques d'enquête*, 14, 257-280.
- MULRY, M.H., et SPENCER, B.D. (1990). Total error in post census and undercount adjustments. *Journal of the American Statistical Association*, 88, 1080-1091.
- MULRY, M.H., et SPENCER, B.D. (1993). Accuracy of the 1990 census of 1988. U.S. Bureau of the Census. *1990 Annual Research Conference Proceedings*, 326-361.
- NATHAN, G. (1967). Outcome probabilities for a record matching process with complete invariant information. *Journal of the American Statistical Association*, 62, 454-469.
- S.I.S. (1994). 1990 Census of Population Response Reliability Survey. *State Institute of Statistics Publication*, No. 1688, Ankara, 65.
- SRINIVASAN, S.K., et MUTTHIAH, S. A. (1968). Problems of matching births identified from two independent sources. *The Journal of Family Welfare*, 14, 13-22.
- TEPPING, B.J. (1968). A model for optimal linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.
- WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.
- GOODMAN, L.A. (1949). On the estimation of the number of classes in a population. *Annals of Mathematical Statistics*, 20, 572-579.
- HARTLEY, H.O. (1962). Multiple frame surveys. Dans *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206
- HARTLEY, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, Series C, 36, 99-118.
- HOGAN, H. (1990). The 1990 Post Enumeration Survey: An Overview. U. S. Bureau of the Census Paper, Washington DC, 6.
- HOGAN, H. (1993a). The 1990 post enumeration survey: operations and results. *Journal of the American Statistical Association*, 88, 1047-1060.
- HOGAN, H. (1993b). Planning for census correction: the 1990 United States experience. Invited Paper, 49th Session of the International Statistical Institute, Florence, Italy. *International Association of Survey Statisticians Booklet*, 133-150.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête post-censitaire. *Techniques d'enquête*, 14, 105-124.
- ISAKI, C.T. (1992). Model bias effects in small area coverage error estimation. *Communication in Statistics Serie A*, 21, 1213-1231.
- MARKS, E.S., SELTZER, W. et KROTKI, K.J. (1974). Population Growth Estimation: A Handbook of Vital Statistics Measurement. New York: The Population Council.

Mr. Ömer Güçelioğlu de leurs commentaires constructifs. Nous remercions aussi Mme Hasibe Dedes et Mme Canan Bakıcı de leur contribution. Enfin, nous exprimons notre reconnaissance au rédacteur, au rédacteur adjoint et à l'examinateur pour leurs commentaires et suggestions. Les opinions exprimées sont celles des auteurs et ne reflètent par forcément celles du State Institute of Statistics, Turquie.

ANNEXES : Outils de dénombrement

Les listes et questionnaires qui suivent sont utilisés avant et durant les opérations du recensement et de l'EP.

Annexe 1. Liste des formules utilisées

Formule 1 : Liste des immeubles (pour les localités dotées

d'une organisation municipale)

Cette liste est créée par le personnel municipal local, puis

produite en trois exemplaires. Elle est utilisée pour la

numérotation séquentielle des UL dans les régions urbaines.

Formule 2 : Liste des immeubles (pour les localités sans

organisation municipale)

Cette liste est créée par le chef de village, puis produite

en trois exemplaires. Elle est utilisée pour la numérotation

séquentielle des unités de logement dans les régions rurales.

Formule C : Liste des secteurs de dénombrement des

immeubles

Cette liste est établie d'après la formule 1 ou 2. Les SD

sont formés d'après cette liste pour les régions urbaines et

rurales séparément.

Formule D : Liste des contrôles du recensement

Il s'agit d'une mise à jour de la formule C qui est remplie

par le recenseur après les opérations de recensement sur le

terrain et remise au comité local de recensement avec les

questionnaires de recensement dûment remplis. Cette

formule et les questionnaires de recensement dûment

remplis sont transmis au SIS après les opérations sur le

terrain.

ANNEXE 2. Questionnaires utilisés

Formule A : Questionnaire du recensement de la

population

Le questionnaire du recensement de la population

comprend quatre grands volets. Les renseignements sont

recueillis durant une interview sur place par la méthode

papier et crayon.

Partie 1. Renseignements sur l'adresse.

Partie 2. Type de lieu de résidence.

Partie 3. Module du ménage

[contiennent sept questions précodées sur le ménage].

L'information est recueillie pour identifier le chef de

ménage et confirmer sa présence, le nombre total de

personnes dans le ménage, le nombre d'invités, le nombre

de membres du ménage absents, la propriété de l'UL

actuelle et la propriété de tout autre UL.

BIBLIOGRAPHIE

Formule B : Questionnaire de l'enquête postcensitaire
En général, le questionnaire de l'EP est basé sur un sous-ensemble de questions de l'étude principale. Cependant, pour la présente étude, le comité consultatif du recensement a décidé d'utiliser le questionnaire complet du recensement pour l'EP. Le questionnaire de l'EP est rempli de la même façon que celui du recensement.

principal.
l'existence d'enfants, la situation d'emploi et l'emploi permanent, le niveau de scolarité, l'état matrimonial, ménage, le lieu de naissance, la citoyenneté, la résidence recueillis sur le sexe, l'âge, la relation avec le chef du Pour chaque personne présente, des renseignements sont [contiennent 26 questions précodées sur les particuliers].

Partie 4. Module des particuliers

CRESSIE, N. (1988). Dans quelles circonstances les opérations de recensement améliorent-elles les chiffres du recensement? *Techniques d'enquête*, 14, 205-222.

DEMING, W.E., et GLASSER, G.J. (1959). On the problem of matching lists by samples. *Journal of the American Statistical Association*, 54, 403-415.

DIFFENDAL, G. (1988). Test des opérations de redressement de 1986 dans le Central Los Angeles County. *Techniques d'enquête*, 14, 75-92.

FAY, R.E., PASSSEL, J.S., ROBINSON, J.G. et COWAN, C.D. (1988). The Coverage of Population in the 1980 Census. *Evaluation and Research Reports*, PHC 80-E4, U.S. Bureau of the Census, 123.

CHANDRA SEKHAR, C., et DEMING, W.E. (1949). On a method of estimating birth and death rates and the extend of registration. *Journal of the American Statistical Association*, 44, 101-115.

CHOI, C.Y., STEEL, D.G. et SKINNER, T.J. (1988). Redressement des chiffres du recensement de 1986 en Australie pour le sous-dénombrement. *Techniques d'enquête*, 14, 187-204.

CASADY, R.J., NATHAN, G. et SIRKEN, M.G. (1985). Alternative dual system network estimators. *Revue Internationale de la Statistique*, 53, 183-197.

BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-79.

AYHAN, H.Ö. (2000). Estimators of vital events in dual-record systems. *Journal of Applied Statistics*, 27, 157-169.

AYHAN, H.Ö., et EKN, S. (1991). Coverage and response errors in 1990 Turkish Census of Population. *Bulletin of the International Statistical Institute*, 54, 45-46.

AYHAN, H.Ö., et EKN, S. (1991). Coverage and response errors in 1990 Turkish Census of Population. *Bulletin of the International Statistical Institute*, 54, 45-46.

Tableau 9
Estimations des taux de couverture du recensement pour les populations régionales et totale des ménages en Turquie en 1990 et leurs erreurs types

h	$\lambda_h^{(1)}$	$se[\lambda_h^{(1)}]$	$\lambda_h^{(2)}$	$se[\lambda_h^{(2)}]$	$\lambda_h^{(3)}$	$se[\lambda_h^{(3)}]$
1	1,14061	0,00410	1,17762	0,00407	1,16127	0,00408
2	1,01463	0,00607	1,05148	0,00607	1,02460	0,00607
3	0,81218	0,00407	0,86803	0,00411	0,82138	0,00408
4	0,95601	0,00718	1,00423	0,00719	0,96293	0,00719
5	0,90272	0,00506	0,96088	0,00508	0,91955	0,00507
Total	0,96466	0,00223	1,01411	0,00223	0,97825	0,00223

recensement ne se manifestent que dans des zones locales très restreintes et les chiffres sont réévalués ultérieurement et éliminés des données du recensement avant la diffusion des résultats.

Selon les résultats de l'étude, il est évident, si l'on s'en tient à la comparaison de plusieurs estimations basées sur des données d'échantillon au dénombrement du recensement, que les méthodes de dénombrement appliquées pour le recensement de la population de la Turquie posent plusieurs problèmes méthodologiques. Les plus importants sont les suivants :

- (1) Améliorer et mettre à jour la liste des SD possibles dans les zones périphériques en expansion rapide des grandes villes en recourant à des méthodes aréolaires.
 - (2) Obtenir une liste parfaite de toutes les unités de logement dans les SD. Le meilleur moyen consisterait à mettre sur pied une opération de dépiage permanent par les autorités locales, qui seraient tenues par la loi de l'exécuter. Alternativement, le SIS pourrait procéder, avant le recensement de la population, à un recensement des logements qui fournirait aussi une base de sondage utile pour le recensement de la population.
 - (3) De nombreux textes législatifs du pays font allusion aux derniers « chiffres de population ». Par conséquent, il pourrait être nécessaire d'apporter des changements importants à ces textes législatifs ainsi qu'aux méthodes de dénombrement.
 - (4) Le dénombrement des populations mobiles exige aussi une attention particulière, l'élaboration de méthodes spéciales et du personnel qualifié.
- Il faut espérer que les fonctionnaires supérieurs responsables de la question considèrent la mesure des caractéristiques de la population grâce au recensement comme une tâche importante et que les progrès nécessaires seront réalisés.

REMERCIEMENTS

Nous remercions les professeurs Oshan Güven, Yalçın Tuncer, Vijay K. Verma, Moti Lal Tikku, M. Qamar Islam et

Un profil net se dégage pour certaines régions, pour toutes les estimations. On peut aussi exprimer les taux de couverture du recensement sous forme de la grandeur de la divergence du recensement. Le profil devrait être le même pour les trois estimateurs, puisque ceux ci sont fortement corrélés.

Pour le total de population, les estimations par les méthodes (1) et (3) donnent un sous dénombrement au recensement comparativement aux chiffres de population réels correspondants. Compte tenu des méthodes de calcul, nous pouvons recommander l'estimateur 3 plutôt que les autres, car il est fondé sur la population projetée des ménages, qui est aussi la population échantillonnée.

Un profil se dégage aussi des résultats pour les estimations régionales quelle que soit la méthode d'estimation. Pour les régions 1 et 2, toutes les estimations témoignent d'un surdénombrement au recensement, tandis que pour toutes les autres régions, elles indiquent un sous dénombrement, sauf pour l'estimateur 2 dans la région 4.

7. CONCLUSIONS

L'étude de l'erreur de couverture du recensement de la population a fourni certains renseignements utiles pour évaluer les problèmes méthodologiques. Ces résultats sont résumés plus bas. La comparaison des trois estimations proposées du chiffre total de population indique que le premier estimateur fournit la valeur la plus élevée de ce total, tandis que l'estimateur 3 donne un résultat plus représentatif de la population totale des ménages.

L'évaluation des taux de couverture du recensement et de la grandeur de la divergence du recensement montrent que, pour la population totale, les estimateurs 1 et 3 produisent un sous dénombrement au recensement. On observe aussi un profil distinct pour les estimations régionales, quelle que soit la méthode d'estimation. Il semble exister un surdénombrement au recensement dans les deux premières régions, mais un sous dénombrement pour les trois autres, dans le cas de toutes les estimations (sauf pour l'estimateur 2 dans la région 4).

Dans les pays en développement, le principal problème que pose le recensement est le sous dénombrement. En Turquie, les problèmes de surdénombrement au

Les estimateurs fondés sur le système d'enregistrement double devraient, en principe, produire des dénombrements estimatifs plus élevés que ceux calculés d'après les données d'une enquête à un seul cycle (c'est-à-dire l'EP). Par conséquent, tous les estimateurs proposés pour le chiffre total de population des ménages sont fondés sur le système d'enregistrement double (SED). Les estimations SED de la population totale des ménages sont présentées au tableau 7.

Les différences entre les trois estimations proposées ne reposent que sur le type de facteur d'expansion utilisé. Si nous examinons ce dernier, $F_h^{(1)}$ est égal au rapport des chiffres projetés de population aux tailles d'échantillon de l'enquête auprès des ménages. Par contre, $F_h^{(2)}$ est égal au nombre total de SD dans l'échantillon correspondant au plan de sondage original de l'EP.

Enfin, $F_h^{(3)}$ est égal au rapport de la taille de la population des ménages à la taille de l'enquête auprès des ménages. Les deux premiers estimateurs incluent la composante des personnes vivant en établissement $[N_h^{(2)}]$ au numérateur de leur facteur d'expansion $[N_h^{(1)}]$ ou $M_h^{(1)}$, tandis que le troisième utilise les données sur la population des ménages $[N_h^{(3)}]$ dans son facteur d'expansion. Manifestement, le facteur d'expansion

Pour pouvoir comparer les statistiques sur l'erreur de couverture, il faut que les chiffres de population soient établis d'après la même base de référence. Il est recommandé d'utiliser un dénombrement de la population des ménages qui correspond à l'estimation démographique. Les chiffres régionaux et total de population sont donnés au tableau 8. Comme nous l'avons mentionné plus haut, les chiffres de la population vivant en établissement sont déterminés d'après les chiffres du Recensement de 1990.

Pour évaluer l'erreur de couverture de la population, nous utilisons le taux de couverture du recensement et total de couverture de la population sont présentés au tableau 9.

6. COMPARAISON DES STATISTIQUES SUR L'ERREUR DE COUVERTURE

du troisième estimateur est calculé d'après les probabilités de sélection idéales $[f_h^{(3)} = n_h^{(1)}/N_h^{(3)}] = 1/F_h^{(3)}$ pour l'échantillon de l'EP, qui est basé sur l'information sur les ménages. Par conséquent, nous pouvons considérer l'estimateur 3 comme étant mieux représentatif de la population des ménages que les autres.

Tableau 7
Estimations des populations régionales et totale des ménages pour 1990 au moyen de l'estimateur fondé sur un système d'enregistrement double étendu et leurs erreurs types

h	$N_h^{(1)}$	$se[N_h^{(1)}]$	$N_h^{(2)}$	$se[N_h^{(2)}]$	$N_h^{(3)}$	$se[N_h^{(3)}]$
1	15 936 939	58 967*	15 436 073	57 113	15 653 378	57 917
2	7 635 314	31 305	7 367 741	30 208	7 561 003	31 000
3	15 573 280	57 621	14 571 298	53 914	15 398 933	56 976
4	6 181 402	38 943	5 884 582	37 073	6 136 969	38 663
5	12 242 987	48 972	11 502 934	46 012	12 019 938	48 080
Total	57 569 922	241 794	54 762 628	230 003	56 770 221	238 435

* : Les estimations de l'erreur-type sont arrondies à l'unité près.

Tableau 8
Chiffres régionaux et total de population pour la Turquie, 1990

h	N_h	$N_h^{(2)}$	N_h^*
Chiffres du recensement	Chiffres de la population vivant en établissement	Chiffres de la population des ménages	
1	18 544 967	367 184	18 177 783
2	7 836 940	89 934	7 747 006
3	12 824 347	176 031	12 648 316
4	5 964 565	55 104	5 909 461
5	11 302 216	249 309	11 052 907
Total	56 473 035	937 562	55 535 473

où $N_h^* = N_h - N_h^{(2)}$

- (5) La comparaison de données sur l'erreur de couverture fondées sur des populations de référence différentes serait incorrecte et conduirait à des résultats fautifs.
- (6) Le sous-dénombrement au recensement est gonflé artificiellement si l'on choisit la population incorrecte (à savoir, la population totale) comme population cible.
- 4.4 Mesures de l'erreur de couverture**
- Nombre de statistiques sur l'erreur de couverture sont proposées dans la littérature. Certaines sont fondées sur de simples ratios ou proportions, et d'autres, sur des méthodes d'ajustement plus complexes. Pour simplifier la résolution du problème, nous proposons d'utiliser comme mesures de l'erreur de couverture pour les populations régionales et totale le taux de couverture du recensement, le taux de divergence du recensement et la divergence du recensement. Ainsi, nous proposons les mesures de l'erreur de couverture qui suivent, qui sont fondées sur la *population des ménages*.

Taux de couverture du recensement :

$$\text{Estimateurs régionaux : } \lambda_{(s)}^h = N_{*}^h / \mathbf{N}_{(s)}^h \quad \forall h = 1, 2, \dots, H \quad (12)$$

où N_{*}^h = Chiffre de population des ménages du recensement [$N_{*}^h = N_{(1)}^h - N_{(2)}^h$] et $\mathbf{N}_{(s)}^h$ = Estimation d'après la source (ou méthode) s .

Erreur-type des estimateurs régionaux : En procédant à la transformation d'échelle $\lambda_{(s)}^h(0,5) = \lambda_{(s)}^h$ qui est une proportion, en tenant compte du fait que, dans chaque strate

l'erreur type du taux de couverture des estimateurs de $\lambda_{(s)}^h + (1 - \lambda_{(s)}^h) = 1$, nous calculons les estimateurs de chaque région comme suit

$$se \left[\tilde{\lambda}_{(s)}^h \right] = \left[\frac{\tilde{\lambda}_{(s)}^h}{1 - \tilde{\lambda}_{(s)}^h} - \frac{n_{(D)}^h}{1 - 1} \right]^{1/2} \quad (13)$$

$$\text{Estimateur du chiffre total de population : } \lambda = N^{*} / \mathbf{N}_{(s)} \quad (14)$$

Taux de divergence du recensement :

$$\text{Estimateurs régionaux : } \phi_{(s)}^h = 1 - \lambda_{(s)}^h = \left[\delta_{(s)}^h / \mathbf{N}_{(s)}^h \right] \quad (15)$$

$$\text{Estimateur du chiffre total de population : } \phi = 1 - \lambda = \left[N^{*} / \mathbf{N}_{(s)} \right] = 1 - \lambda_{(s)} \quad (16)$$

Divergence du recensement :

$$\text{Estimateurs régionaux : } \delta_{(s)}^h = \mathbf{N}_{(s)}^h - N_{*}^h \quad \forall h \quad (17)$$

Etant donné les limites d'un recensement en un jour par le système *de facto*, nous n'avons pu considérer d'autres mesures de la couverture locale pour la présente étude. Des

5. ESTIMATEURS DU TOTAL DE POPULATION

Le total de population estimé correspond à la somme pondérée des estimations régionales.

$$\mathbf{N}_{(s)}^h = \sum_H n_{(s)}^h \quad (18)$$

Les estimateurs de l'erreur type de la population totale des ménages de chaque région sont donnés par

$$se \left[\mathbf{N}_{(s)}^h \right] = \mathbf{N}_{(s)}^h \left[\frac{p_{(D)}^h (1 - p_{(D)}^h)}{1 - 1} \right]^{1/2} \quad (19)$$

tandis que la proportion imputable à chaque strate est donnée par $p_{(D)}^h = n_{(D)}^h / \sum_H n_{(D)}^h$.

Le calcul de l'erreur de couverture pour un recensement particulier n'est pas une tâche facile, surtout si l'on ne dispose pas d'une liste parfaite de la population cible pour comparer les résultats. Il en est ainsi dans la plupart des pays, sauf ceux qui possèdent des registres de population. La comparaison des résultats d'un recensement de population aux projections démographiques pose aussi certains problèmes, liés à la validité des hypothèses qui sous-tendent les modèles de projection. Pour éviter de fonder des comparaisons sur une seule base, nous proposons les *estimateurs régionaux fondés sur un système d'erreur-gistivement double élargis* pour la détermination de l'erreur de couverture du recensement.

$$\text{Estimateur 1. } \mathbf{N}_{(1)}^h = F_{(1)}^h(n_{(D)}^h) \quad (20)$$

$$\text{où } F_{(1)}^h = N_{(1)}^h / n_{(1)}^h \text{ et } n_{(D)}^h = \sum_c \sum_r n_{(r,c)}^h \quad (21)$$

Ici, $n_{(D)}^h$ représente l'estimation SED non pondérée et $\mathbf{N}_{(1)}^h$ correspond à la taille de l'échantillon sélectionné.

$$\text{Estimateur 2. } \mathbf{N}_{(2)}^h = F_{(2)}^h(n_{(D)}^h) \text{ où } F_{(2)}^h = M_{(1)}^h / m_{(1)}^h \quad (22)$$

$$\text{Estimateur 3. } \mathbf{N}_{(3)}^h = F_{(3)}^h(n_{(D)}^h) \text{ où } F_{(3)}^h = N_{(3)}^h / n_{(1)}^h \quad (22)$$

Tableau 5

Nombre de personnes appartées et non appartées pour le recensement et l'enquête postcensitaire, selon la région

Régions	Appartées	Recensement Non appartées	EP Non appartées	Nbre estimé d'omissions dans les deux sources	Estimation basée sur le système d'entreg. double
h	$n_h(1, 1)$	$n_h(1, 2)$	$n_h(2, 1)$	$n_h(2, 2)$	$n_h^{(D)}$
1	13 393	642	533	26	14 594
2	6 400	187	182	5	6 774
3	11 644	1 414	1 340	163	14 561
4	3 134	1 099	446	156	4 835
5	6 449	1 449	1 439	323	9 660
Total	41 020	4 791	3 940	673	50 424

Tableau 6

Détermination de la taille de la population des ménages et de l'échantillon selon la région

Régions	Taille de population prévue	Estimation de la population vivant en établissement	Taille de la population des ménages	Taille de l'échantillon de ménages	Facteurs d'expansion
h	$N_h^{(1)}$	$N_h^{(2)}$	$N_h^{(3)}$	$n_h^{(1)}$	$F_h^{(1)}$
1	20 639 200	367 184	20 272 016	18 900	1 092,02
2	9 242 600	89 934	9 152 666	8 200	1 127,15
3	15 731 600	176 031	15 555 569	14 800	1 062,95
4	7 670 800	55 104	7 615 696	6 000	1 278,47
5	13 687 800	249 309	13 438 491	10 800	1 267,39
Total	66 972 000	937 562	66 034 438	58 700	1 140,92
					$F_h^{(3)}$
					1 124,95

4.3 Chiffre total de population contre population des ménages

Pour les projections démographiques utilisées en vue d'estimer le nombre total de SD dans la population, on a considéré le chiffre total de population comme étant la population cible. Par contre, dans le plan d'échantillonnage de l'EP, la population cible est limitée à la population des ménages, parce que le plan est fondé uniquement sur les unités de logement échantillonnées, qui excluent les secteurs de dénombrement spéciaux (population vivant en établissement).

Comme nous l'avons mentionné plus haut, le plan d'échantillonnage de l'EP a servi de base pour la comparaison des deux systèmes distincts de dénombrement durant la procédure d'appariement. Il a donc été naturel de considérer la population des ménages comme étant la population cible pour l'estimation appropriée du chiffre total de population au moyen des estimateurs proposés. À cette fin, nous avons déterminé ultérieurement le chiffre de la population vivant en établissement à partir des données de Recensement de 1990, pour les régions et les tranches de taille de population. La population vivant en établissement est présentée par région (en agrégeant sur les tranches de

Lors de l'utilisation subséquente des données sur la population vivant en établissement, nous avons également émis l'hypothèse qu'aucune erreur de couverture n'était associée à la mesure de cette population faite lors du Recensement de 1990. Puis, pour chaque région, nous avons calculé la population des ménages par différence. Les raisons de soustraire la population vivant en établissement du chiffre total de population sont multiples. Le plan d'échantillonnage de l'EP ne reflète que la population des ménages.

(1) Les probabilités de sélection correctes pour la couverture (représentation) idéale pour chaque strate de l'échantillon doivent être fondées sur la population des ménages et non sur le total de la population.

(2) Les estimations proposées de l'erreur de couverture ne devraient se fonder que sur la population des ménages.

(4) Les estimateurs proposés du chiffre total de population devraient également se fonder sur la population des ménages, quand les résultats de l'EP sont fondés sur les ménages.

recommandé, pour le dénombrement de ces dernières, d'adopter le système de jure plutôt que le système de facto, et de recenser les recenseurs mobiles plutôt que stationnaires.) parce qu aucune liste n existe.

(11) Les SD du recensement et ceux de l'EP sont d'immuables sont établies par les autorités locales et ne sont pas suffisamment fiables pour certains recensements. La numération des SD est également réalisée au niveau local et est également affectée par le manque de fiabilité des opérations de numération.

4.2 Estimation fondée sur un système d'enregistrement double

On utilise la méthode du système d'enregistrement double pour estimer les nombres de ménages et de particuliers par une méthode d'appariement. Les résultats ainsi obtenus sont utilisés pour estimer le nombre total de personnes dans chaque région ainsi que le chiffre total de population. Le modèle suppose l'indépendance de la collecte des données provenant des deux sources, ici le recensement et l'EP. Théoriquement, toutes les cellules $[n(r, c)]$ sont observables, sauf pour $n(2, 2)$ et tout total qui inclut $n(2, 2)$. Chandra Sekar et Deming (1949) postulent qu'il n'existe aucun biais de corrélation dans l'estimation pour la cellule $n(2, 2)$. Pour des raisons pratiques, nous supposons ici que cette hypothèse est valide. Une discussion plus approfondie de la validité de ce genre d'hypothèse a été publiée récemment par Ayhan (2000).

La méthodologie et les procédures d'estimation sont décrites ci après. L'estimation du nombre de personnes non dénombrées lors du recensement ou de l'EP est donnée par :

$$(9) \quad n(2, 2) = [n(1, 2) \cdot n(2, 1)] / n(1, 1).$$

Le nombre total de personnes est estimé comme suit :

$$(10) \quad n = n(1, 1) + n(1, 2) + n(2, 1) + n(2, 2)$$

ou bien comme suit,

$$(11) \quad n = [n(*, 1) \cdot n(1, *)] / n(1, 1).$$

Le tableau 2 présente plus haut illustre la méthode d'appariement utilisée pour le système d'enregistrement double. La méthode de calcul décrite ici a été répétée séparément pour chaque région. Les estimations sont données au tableau 5. Pour chaque strate, n_h est calculé comme indiqué plus haut pour n .

Les raisons des différences de couverture des SD lors du recensement et de l'EP de la Turquie sont les suivantes.

- (1) Des formules C et D supplémentaires sont établies par le comité du recensement des provinces d'après la liste des immuables (formules 1 et 2). Les listes d'immuables sont établies par les autorités locales et ne sont pas suffisamment fiables pour certains recensements.

- (2) La numération des SD est également réalisée au niveau local et est également affectée par le manque de fiabilité des opérations de numération.
- (3) Les formules C et D peuvent ne pas contenir 100 personnes pour les régions urbaines et 200 personnes pour les régions rurales, parce que certaines listes ne sont pas à jour.

- (4) À cause de points de départ différents, les nombres d'unités de logement couverts par les recenseurs et par les enquêteurs de l'EP sont différents.

- (5) L'administration du questionnaire de l'EP a débuté au moins deux heures après le début des opérations du recensement dans les SD échantillonnées. Les différences de couverture pourraient être attribuables à la mobilité des membres des ménages recensés dans le même SD.

- (6) Durant la période de recensement d'un jour, certains questionnaires prévus du recensement et de l'EP n'ont pu être remplis, ce qui a produit des incohérences durant l'appariement. Il s'agit naturellement d'une source de sous dénombrement, mais la situation se produit rarement durant les opérations sur le terrain.

- (7) Comme le recensement est fondé sur le système de facto, les visiteurs locaux (provenant d'autres logements de l'immuable) peuvent donner lieu à des changements pour chacune des sources de données. De nouveau, à cause du système de dénombrement de facto, des erreurs de dénombrement peuvent se produire pour la population mobile du recensement. L'EP ne couvre que la population des ménages.

- (9) L'EP n'est pas conçue pour couvrir les SD spéciaux ni les populations mobiles (c'est-à-dire les voyageurs, les personnes de garde, etc.). Par définition, les voyageurs étrangers et nationaux ont le droit de poursuivre leur voyage après avoir été dénombrés, si leur voyage a débuté avant l'heure officielle de début du recensement. Pour la présente étude, la population mobile a été exclue de l'analyse.

- (10) L'EP ne couvre par les tribus nomades (Le recensement des tribus nomades nécessite le recours à des techniques de dénombrement spéciales. Il est

Cliché de la méthode d'appariement

Tableau 2

Méthode d'appariement			
DONNÉES 2 : (EP)			
SOURCE DE	Dedans	Pas dedans	Total
DONNÉES 1 :			
DE			
Dedans	$n(1, 1)$	$n(1, 2)$	$n(1, *)$
Pas dedans	$n(2, 1)$	$n(2, 2)$	$n(2, *)$
Total	$n(*, 1)$	$n(*, 2)$	n

Conformément aux spécifications susmentionnées, à la première étape, on apparie les ménages et, à la deuxième

étape, on apparie les personnes comprises dans les ménages appariés. Les résultats sont présentés dans les tableaux qui suivent par région. Les secteurs de dénombrement sont situés dans les peuplements échantillonnés dans 19 provinces réparties entre cinq régions du plan d'échantillonnage. Sur les 443 SD sélectionnés, 437 ont été appariés à leur homologue du recensement de la population et 6 n'ont pu l'être à cause de différences dans les instructions données indépendamment pour leur création par les bureaux locaux. Les renseignements sur la ventilation régionale des six SD non appariés sont donnés au tableau 3, mais ceux sur la ventilation urbaine-rurale n'ont pas été obtenus. La méthode d'appariement des ménages peut être représentée par $k(r, c)$ de la même façon que celle pré-sentée pour les particuliers au tableau 3. Dans le cas stratifié, on peut aussi représenter le nombre de ménages dans chaque strate par $k_h^*(r, c)$. Le nombre total de ménages du recensement qui ne sont pas appariés à des ménages de l'EP est estimé, pour chaque strate, par

$$k_h^*(2, 1) = [k_h^*(*, 1) - k_h^*(1, 1)] \quad (8)$$

et le nombre total de ménages de l'EP qui ne sont pas appariés à des ménages du recensement est estimé, pour chaque strate, par

$$k_h^*(1, 2) = [k_h^*(1, *) - k_h^*(1, 1)] \quad (7)$$

L'EP est estimé, pour chaque strate, par le recensement qui ne sont pas appariés à des ménages de l'EP est estimé, pour chaque strate, par

Ménages appariés et non appariés des secteurs de dénombrement de l'enquête postcensitaire et du recensement, par région

Tableau 3

Régions	Nbre de SD	Nbre de SD échantillonnées	SD recensées	Ménages appariés	$m_h^{(1)}$	$m_h^{(2)}$	$k_h(1, 1)$	$k_h(1, 2)$	$k_h(2, 1)$	$k_h(2, 2)$
1	157	154	3 320	168	144	27	30	262	259	80
2	62	62	1 262	168	144	27	30	262	259	80
3	112	112	2 636	262	259	80	175	170	175	175
4	38	38	645	204	80	175	175	170	175	175
5	74	71	995	170	175	175	175	170	175	175
Total	443	437	8 858	831	688	175	175	170	175	175

Tableau 4
Nombre de ménages et de particuliers couverts par le recensement et par l'enquête postcensitaire, par région

Régions	Recens.	EP	Appariés	Couverture
1	3 488	3 464	1 0069	1 0078
2	1 289	1 292	0,9977	1,0008
3	2 898	2 895	1,001	1,0057
4	849	725	1,171	1,1824
5	1 165	1 170	0,9957	1,0013
Total	9 689	9 546	1,015	1,0189

Taux de couverture : $C_h = k_h(1, *) / k_h^*(1, *)$ et $C_h^* = n(1, *) / n(*, 1)$

Les données sur les ménages appariés et non appariés sont présentées au tableau 3.

Dans ces 437 SD, 8 858 ménages ont été appariés. En tout, 831 ménages du recensement (9,38 %) et 688 ménages de l'EP (7,77 %) n'ont pu être appariés. Le taux d'appariement des ménages fondés sur le recensement est de 90,62 %, tandis que le taux d'appariement des ménages fondés sur l'EP est de 92,23 %, d'après les données présentées au tableau 3.

Les taux de couverture des ménages pour le recensement et l'EP sont présentés par région au tableau 4. Pour la plupart des régions (sauf les régions 2 et 5) et pour l'ensemble de celles-ci, le taux de couverture des ménages (C_h^*) est plus élevé dans le cas du recensement que dans celui de l'enquête postcensitaire. Ici, tous les taux de couverture sont plus élevés que prévu. En ce qui concerne les particuliers faisant partie des ménages couverts, le taux de couverture (C_h^*) est plus élevé dans le cas du recensement, pour toutes les régions et pour l'ensemble de celles-ci. Le nombre total de particuliers appariés est $n(1, 1) = 41 020$ pour le recensement et pour l'EP.

4. MÉTHODES D'ESTIMATION DE L'ERREUR

DE COUVERTURE

Nous abordons maintenant l'estimation de l'erreur de couverture en décrivant les méthodes d'appariement des données, les méthodes d'estimation fondées sur un système d'enregistrement double et les résultats obtenus. On se sert, pour estimer la couverture de la population, de listes de SD provenant de deux sources indépendantes. Dans la présente section, nous présentons les méthodes d'appariement des données, les estimateurs fondés sur un système d'enregistrement double et d'autres estimateurs du total de population, et nous évaluons les estimations obtenues. Nous comparons aussi les statistiques sur les erreurs de couverture calculées.

4.1 Méthodes d'appariement des données

Plusieurs modèles (Deming et Glasser 1959; Nathan 1967 et Tepping 1968) ont été proposés pour déterminer les méthodes optimales d'appariement. Ils s'appuient sur l'état-blessement de procédures qui réduisent au minimum l'erreur d'appariement nette estimée sous réserve des coûts et autres contraintes (Marks, Seitzer et Krotki 1974). Ces modèles offrent des concepts d'appariement valables en théorie et en pratique, mais aucun n'est entièrement polyvalent. Les travaux de Tepping (1968), étendus par Srinivasan et Muthiah (1968), s'appuyaient sur un ensemble minimal de caractéristiques pour la concordance exacte lors de l'appariement. En outre, Ayhan et Eknı (1991) et le SIS (1994) ont utilisé des méthodes comparables, fondées sur les spécifications qui suivent.

(1) *Appariement de la population des SD*. La population totale d'un SD est égale à la somme de la population des ménages de toutes les unités de logement (UL) comprises dans le SD.

(2) *Appariement des ménages à l'intérieur des SD*. Plusieurs ensembles d'information (adresse de l'unité de logement, nom du chef de ménage et nombre de personnes dans le ménage) sont utilisés pour procéder à l'appariement des ménages.

(3) *Appariement des particuliers à l'intérieur des ménages*. En tout, quatre variables du recensement de l'EP (nom, âge, sexe et niveau de scolarité) ont été utilisées pour établir la concordance exacte lors de l'appariement des particuliers.

(4) *Appariement des particuliers non apparés des ménages*. Cette tâche est réalisée par appariement aux particuliers faisant partie des ménages voisins (provenant de l'autre source de données), par recherche. L'appariement des particuliers se fonde sur les mêmes critères de concordance exacte.

Les travaux préliminaires des opérations d'appariement sont réalisés manuellement, tandis que l'évaluation des appariements de ménages et de particuliers est automatisée. Le tableau 2 donne les fréquences $n(r, c)$ pour la méthode d'appariement des particuliers.

On procède à l'échantillonnage systématique de 443 SD

d'après la liste ainsi établie.

Les fractions d'échantillonnage et la répartition de l'échantillon sont obtenues de la façon suivante. Les secteurs de dénombrement de l'échantillon sont sélectionnés dans toutes les strates par échantillonnage avec probabilités égales. La fraction d'échantillonnage prévue est $f_h \approx 0,001$ pour toutes les strates. Toutefois, elle varie d'une strate à l'autre. Suivent des renseignements techniques sur les fractions d'échantillonnage et la répartition de l'échantillon $[n_h^{(c)}]$ s'obtiennent comme suit.

$$f_{(1)}^h = n_{(1)}^h / N_{(1)}^h = 1 / F_{(1)}^h \text{ et } f_{(2)}^h = m_{(1)}^h / M_{(1)}^h = 1 / F_{(2)}^h. \quad (1)$$

Les tailles totales de population des SD urbains (U) et ruraux (R) sont

$$N_{(U)}^h = M_{(U)}^h B_{hi} = \sum_{i=1}^I M_{hi} B_{hi} \quad \forall h \& i \quad (2)$$

$$N_{(R)}^h = M_{(R)}^h B_{hj} = \sum_{j=1}^J M_{hj} B_{hj} \quad \forall h \& j \quad (3)$$

où les composantes ont été définies antérieurement. Puis, la taille de population de chaque strate est déterminée comme suit

$$N_{(1)}^h = [N_{(U)}^h + N_{(R)}^h]. \quad (4)$$

De la même façon, la taille d'échantillon correspondante de chaque strate est :

$$n_{(1)}^h = [n_{(U)}^h + n_{(R)}^h] \quad (5)$$

$$\text{où } n_{(U)}^h = m_{(U)}^h B_{hi} \text{ et } n_{(R)}^h = m_{(R)}^h B_{hj} \quad (6)$$

3.2 Conception des opérations de l'EP

Les opérations sur le terrain de l'EP sont identiques à celles du recensement, pour lesquelles des renseignements sont données à la section 2.2. Pour des raisons opérationnelles, chaque SD est défini comme étant un intervalle fermé de nombres d'unité de logement dans les rues. En ce qui concerne les secteurs de dénombrement spéciaux (c'est-à-dire établisements), le nombre total est vérifié d'après des renseignements antérieurs obtenus au niveau provincial. Conformément aux instructions données aux recenseurs, l'EP débute dans les secteurs de dénombrement de l'échantillon une heure après le début du recensement, le même jour que celui-ci. Les enquêteurs de l'EP rendent visite aux ménages de sorte qu'ils ne prennent contact avec aucun ménage avant que les recenseurs ne l'aient fait. Les résultats de l'EP sont utilisés pour l'évaluation, après appariement des cas individuels aux enregistrements du recensement pour les SD correspondants.

(4) Des erreurs de traitement, comme celles commises par les codeurs et les vérificateurs des données, surviennent aussi durant le traitement des données et sont éliminées ultérieurement lors du traitement des données au bureau central.

3. PROCÉDURES DE L'ENQUÊTE POSTCENSITAIRE

Les objectifs de l'EP consistent à évaluer l'erreur de couverture du recensement de la population et à obtenir des mesures de la fiabilité des réponses aux questions du recensement. Dans le présent article, nous discutons du premier objectif dans le cas du Recensement de la population de la Turquie. Les observations préliminaires concernant les deux objectifs sont résumées dans Ayhan et Eknî (1991).

3.1 Méthode de sélection de l'échantillon

L'établissement du plan d'échantillonnage de l'EP débute trois mois avant les opérations de recensement. À ce stade, la création des secteurs de dénombrement (SD) du recensement n'est pas encore terminée.

Stratification et estimation de la population des SD. Les listes de secteurs de dénombrement établies par le State Institute of Statistics pour le recensement de la population précédent servent de base d'échantillonnage pour les opérations de l'EP. La population est d'abord stratifiée en cinq *régions géographiques-socioéconomiques* de la Turquie. On se sert aussi d'une deuxième variable de stratification explicite, qui est basée sur les huit tranches de taille selon *le lieu du peuplement*, correspondant à une structure hétéroclique à l'intérieur des régions. Ici, la limite urbaine-rurale correspond à une taille de population de 10 000. Le nombre de secteurs de dénombrement est estimé pour 40 strates du plan de sondage, le jour du recensement, par une méthode de projection démographique, basée sur les dénombrements des deux recensements de la population précédents. Pour le recensement, les SD sont créés au moyen de la formule C par le bureau central. En tout, 479 251 SD ont été établis pour le dernier recensement. Les renseignements sur la base de sondage figurent au tableau 1.

La couverture du nombre d'unités de logement par le recensement et par l'EP est établie de la façon suivante. Dans chaque province, on détermine le nombre de SD dans la population et on les numérote séquentiellement. Puis, pour chaque strate, on estime le nombre de SD dans la population en divisant la population projetée de la strate (N_h) par la charge de travail quotidienne fixe des recenseurs (B_h) . Le nombre de SD dans la population est estimé pour les régions urbaines comme étant $M_{hi} = N_{hi} / B_{hi}$ et pour les régions rurales comme étant $M_{hj} = N_{hj} / B_{hj}$, où les tailles des SD correspondent aux charges de travail quotidiennes fixes, à savoir $B_{hi} = 100$ personnes dans les strates urbaines et $B_{hj} = 200$ personnes dans les strates

rurales. On obtient aussi les projections démographiques pour chaque strate selon l'agrégation urbaine-rurale. Enfin, on calcule le nombre estimatif de SD dans la population et les facteurs d'expansion pour les régions et les strates urbaines-rurales.

Tableau 1

Nombre estimés de SD dans la population et dans l'échantillon selon la région et la strate urbaine-rurale

Région	SD pop.	SD	URBAINE	SD pop.	SD	RURALE	SD pop.	SD	TOTAL
h	$M_{(U)}^h$	$m_{(U)}^h$	$M_{(R)}^h$	$m_{(R)}^h$	$M_{(1)}^h$	$m_{(1)}^h$	$M_{(1)}^h$	$m_{(1)}^h$	
1	125 726	125	40 333	32	166 059	157			
2	42 442	42	24 992	20	67 434	62			
3	65 466	76	45 925	36	111 391	112			
4	15 790	16	30 459	22	46 249	38			
5	39 358	40	48 760	34	88 118	74			
Total	288 782	299	190 469	144	479 251	443			

Facteurs d'expansion : $F_{(1)}^h = (N_{(1)}^h / m_{(1)}^h) \neq M_{(1)}^h / m_{(1)}^h = F_{(2)}^h$

Sélection des SD de l'échantillon. Pour l'EP, on sélectionne systématiquement un échantillon stratifié à plusieurs degrés de localités et d'îlots à partir de la base de sondage principale du State Institute of Statistics tenue à jour par le bureau central. La liste des îlots de la base de sondage principale est mise à jour périodiquement en vue de la sélection polyvalente régulière d'autres échantillons. Le bureau central recrute les intervieweurs de l'EP, assure leur formation, puis les envoie en équipes dans les peuplements locaux échantillonnés où ils procèdent au dénombrement indépendant de l'échantillon de l'EP. Aux fins d'identification, les îlots échantillonnés sont reliés sur le terrain aux SD du recensement de la population correspondants du peuplement, selon des instructions préalables données aux intervieweurs de l'EP.

Pour que l'estimation fondée sur un système d'enregistrement double soit valide, il faut que les secteurs de dénombrement de l'échantillon de l'EP soient déterminés indépendamment de la base de recensement. Il s'agit d'une hypothèse absolument cruciale du modèle SED sur laquelle nombre de chercheurs ont insisté au cours des 50 dernières années (Ayhan 2000 Chandra Sekar et Deming 1949). Etant donné l'utilisation de listes comptant d'anciens SD non soustraites dans certaines régions, l'importance de la charge de travail prévue par SD par intervieweur peut changer et, par conséquent, la taille prévue des SD peut différer de la taille effectivement dénombrée, ce qui influe sur la fraction d'échantillonnage réalisée qui diffèrera évidemment de celle choisie.

En tout, $m = 443$ secteurs de dénombrement d'échantillon ont été sélectionnés dans 16 provinces, 23 districts, 16 sous districts et 43 villages dans les 40 strates. Pour l'EP,

d'heures de prolongation, durant la même journée de recensement.

Des recenseurs supplémentaires sont désignés pour les lieux où se situent les personnes mobiles (voyageurs, personnes de garde, tribus nomades, etc.) et les personnes vivant en établissement (hôpitaux, prisons, usines, établissements militaires, etc.).

La population en établissement est couverte par des recenseurs supplémentaires qui sont affectés aux SD spéciaux. Les personnes mobiles se déplaçant en véhicule sont artées et dénombrées en tant que groupes au moment où elles apparaissent pour la première fois sur le territoire d'une des provinces. Les passagers poursuivent leur voyage après le dénombrement, et un signe indiquant *recensé* est appliqué sur le véhicule après l'opération de recensement pour éviter les doubles comptes; plus tard, l'identité des individus est vérifiée au moyen d'algorithmes manuels ou informatisés, en regard des autres enregistrements obtenus pour le peuplement pertinent. Le recensement a été réalisé pour le dimanche et les travaux de dénombrement ont été achevés ce jour-là. Le jour du recensement, le gouvernement a déclaré un couvre feu national. Les recenseurs ont rendu visite à chaque ménage (M) occupant les unités de logement figurant sur la liste d'immeubles de leur secteur de dénombrement et ont rempli le questionnaire du recensement (voir l'annexe 2 pour des précisions). Pour le *module du ménage*, les renseignements sur les caractéristiques générales du ménage sont recueillis auprès d'un membre adulte du ménage, tandis que pour le *module des particuliers*, les renseignements sur les caractéristiques personnelles sont recueillis auprès des personnes proprement dites.

Suit la liste des erreurs survenues aux diverses étapes des opérations du recensement.

(1) Des erreurs d'omission et d'inclusion ont eu lieu durant l'établissement de la liste d'immeubles. Cependant, grâce à l'utilisation de la méthode du segment indivisible durant le processus de dénombrement du recensement, ces erreurs sont en grande partie éliminées.

(2) Durant les opérations de dénombrement, il se produit des erreurs de réponse dues à des problèmes de ré-mémorisation, au trichage ou à des réponses impossibles à coder. Ces erreurs sont mesurées en tant qu'incohérence des réponses lors de l'étude de la *fiabilité de la réponse* au Recensement de la population (Ayhan et Eknı 1991; SIS 1994) qui est fondée sur l'EP.

(3) Certaines erreurs commises par les recenseurs (omissions des questions d'approfondissement, interprétation inadéquate de la réponse et erreurs d'enregistrement) sont également observées durant les opérations de recensement. Ces erreurs sont, elles aussi, couvertes par l'étude de la *fiabilité de la réponse*.

de facto est d'un usage très répandu dans les pays en voie de développement, le système *de jure* est celui généralement utilisé par les pays développés. Les pays qui utilisent le système de recensement *de facto* semblent avoir plus de problèmes de couverture que ceux qui utilisent le système *de jure*. Ces problèmes tiennent principalement aux imperfections des bases de sondage existantes de la population cible.

Habituellement, les recensements de population fondés sur le système *de facto* sont réalisés en un seul jour, sous forme de dénombrement complet, pour déterminer le chiffre total de population le jour du recensement. Les citoyens du pays vivant à l'étranger sont exclus du recensement, tandis que les étrangers présents dans le pays y sont inclus.

Conception des opérations de recensement

Le recensement de la population de la Turquie est réalisé, selon le système *de facto*, par le State Institute of Statistics (SIS) pour déterminer les caractéristiques quantitatives, sociales et économiques de la population. Aux fins du recensement, les autorités locales dressent la liste des immeubles et la transmettent aux comités locaux du recensement. À partir de cette liste des immeubles (formule 1 ou 2), on établit la liste des *secteurs de dénombrement (SD) des immeubles* (formule C) (voir l'annexe 1 pour des précisions). Comme les listes d'immeubles ne sont pas transmises suffisamment tôt au bureau central du SIS, celui-ci estime le nombre de SD par projection pour planifier les opérations sur le terrain du recensement et de l'EP. Pour former les SD, on affecte 100 personnes par recenseur dans le cas des provinces et des districts, et 200 personnes par recenseur dans le cas des sous-districts et des villages, en fonction des charges de travail quotidiennes moyennes. Puis, les SD sont numérotés séquentiellement. Dans le cas du recensement, les adresses figurant sur les listes sont utilisées pour déterminer les « unités de logement (UL) », tandis que les « particuliers » compris dans le ou les ménages occupant l'unité de logement sont considérés comme étant l'unité de dénombrement.

La charge de travail de chaque recenseur correspond à un SD, qui correspond à une liste d'adresses qui doivent être couvertes dans un intervalle de temps fixé. Les recenseurs ont l'instruction de traiter cet intervalle comme un segment indivisible. Ceux qui découvrent des adresses ne figurant pas sur la liste les incluent dans le recensement conformément à la définition de leur tâche. Dans le cas d'unités inoccupées ou inexistantes, l'information pertinente est également consignée. Aucune procédure particulière n'est établie pour traiter le cas des personnes qui hésitent à répondre ou, de façon générale, toute unité non interviewée, étant donné la nature obligatoire de la participation en vertu de la loi pertinente. La charge de travail des recenseurs est établie de façon à ce qu'ils puissent achever toutes les interviews en une journée. Dans des cas très spéciaux, ils ont l'instruction d'achever le recensement du segment au cours

Erreur de couverture des recensements de population : Le cas de la Turquie

H. ÖZTAŞ AYHAN et SÜHENDAN EKİN¹

RÉSUMÉ

Les erreurs et d'autres problèmes de couverture associés aux recensements de population sont examinés à la lumière des travaux publiés récemment. Plus précisément, quand on apparie les dénombremments réels du recensement aux chiffres correspondants tirés de l'enquête postcensitaire, on obtient des résultats agréables fondés sur un système d'enregistrement double qui fournissent certaines statistiques sur l'erreur de couverture.

Dans le présent article, les questions liées à l'erreur de couverture et diverses solutions sont examinées dans le contexte des résultats du dernier Recensement de la Turquie. La comparaison, au niveau régional, de la couverture du recensement fondée sur les données de ce dernier et celles de l'enquête postcensitaire témoigne d'une variabilité interrégionale. Certaines recommandations méthodologiques sont faites en vue d'une amélioration éventuelle des procédures courantes de dénombrement.

MOTS CLÉS : Erreur de couverture du recensement; mesures de l'erreur de couverture; estimation de l'erreur de couverture; estimation fondée sur un système d'enregistrement double; recensement de la population; enquête postcensitaire.

1. INTRODUCTION

La question de la couverture est importante aussi bien dans le cas des recensements que dans celui des enquêtes par sondage. L'écart entre le chiffre de recensement et le chiffre de population cible représente l'erreur de couverture. Si le chiffre de recensement est inférieur au chiffre de population cible, il y a sous-dénombrement, phénomène courant dans nombre de pays.

Plusieurs méthodes permettent d'étudier le problème des erreurs de couverture lors des recensements. L'estimateur fondé sur un système d'enregistrement double ou SED (Chandra Sekar et Deming 1949) a été étendu par de nombreux chercheurs (Ayhan 2000; Casady, Nathan et Sirken 1985; Hogan 1990, 1993a et 1993b; Isaki 1992; Marks, Seltzer et Krotki 1974).

Le U.S. Census Bureau utilise les estimations à système d'enregistrement double fondées sur le recensement et une enquête postcensitaire (EP) pour évaluer l'erreur de couverture du recensement (Hogan 1993a et 1993b, Mulry et Spencer 1988, 1990 et 1993). La réalisation d'enquêtes postcensitaires permet d'améliorer les estimations démographiques (Ayhan et Ekin 1991; Diffendal 1988; Hogan 1990; Hogan et Wolter 1988).

Aux États Unis, dans le cadre du programme de l'enquête postcensitaire de 1980, on s'est efforcé de mesurer directement la couverture du recensement au moyen de modèles d'enquête par sondage (Fay et coll. 1988). Plusieurs méthodes ont également été proposées pour corriger les chiffres de recensement pour le sous-dénombrement (Choi, Steel et Skinner 1988; Cressie 1986 et 1990).

Récemment, les modèles d'erreur de couverture de la population ont fait l'objet d'études importantes (Isaki 1992; Wolter 1986). Certains chercheurs ont utilisé une méthode basée sur des systèmes de données qui se chevauchent ou sur des bases de sondage multiples pour améliorer les estimations démographiques (Goodman 1949; Hartley 1962 et 1974; Bankier 1986).

La présente étude souligne les problèmes méthodologiques que pose l'évaluation de la couverture des recensements de population et offre des moyens de remédier à certains problèmes décrits. Sont aussi proposées et discutées d'autres estimations des erreurs de couverture du recensement de la population. Pour réaliser ces objectifs, on a tenu compte des problèmes d'évaluation de la couverture lors de la conception de l'EP.

La présentation de l'article est la suivante. La section 2 est consacrée à la discussion des méthodes de dénombrement et la section 3, aux procédures de l'enquête postcensitaire. La section 4 présente les méthodes d'estimation de l'erreur de couverture. La section 5 donne les estimateurs du chiffre de population et la section 6, une comparaison des données sur l'erreur de couverture. Les résultats importants sont résumés dans la conclusion.

2. MÉTHODES DE RECENSEMENT

Les recensements de population de la plupart des pays présentent de nombreuses caractéristiques communes. La méthode de recensement peut être fondée sur un système de *facto* ou de *jure*. Dans le cas du système de *jure*, les personnes sont recensées à leur lieu de résidence ordinaire, tandis que dans le système de *facto*, elles le sont à l'endroit où elles se trouvent effectivement. Alors que le système

KOSTANICH, D. (2001). Accuracy and Coverage Evaluation Survey: computer specifications for Person Dual System Estimation (U.S.) - Re-issue of Q-29. *DSSD Census 2000 Procedures and Operations Memorandum Series* Q-37.

NAVARRO, A., et OLSON, D. (2001). Accuracy and Coverage Evaluation: effect of targeted extended search. *DSSD Census 2000 Procedures and Operations Memorandum Series* B-18*.

PETERSON, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station*, 6, 1-48.

PETRONI, R. (1997). Effect of using the 1996 ICM characteristic imputation and probability modeling methodology on the 1995 ICM P and E-sample data. *DSSD Census 2000 Dress Rehearsal Memorandum Series*, A-20.

PETRONI, R. (1998a). Effect of different methods for calculating match and residence probabilities for the 1995 P-sample data. *DSSD Census 2000 Dress Rehearsal Memorandum Series*, A-23.

PETRONI, R. (1998b). Effect of different methods for calculating correct enumeration probabilities for the 1995 E-sample data. *DSSD Census 2000 Dress Rehearsal Memorandum Series*, A-28.

PETRONI, R. (1998c). Effect of using simple ratio methods to calculate P-sample residence probabilities and E-sample correct enumeration probabilities for the 1995 data. *DSSD Census 2000 Dress Rehearsal Memorandum Series*, A-30.

PETRONI, R. (2001). EFU Sample Design, Stratification, Selection, and Weighting. Planning, Research, and Evaluation Division *TXE/2010 Memorandum Series*. CM-GE5-S-02-R2.

RUBIN, D.B. (1976). Inference and Missing Data. *Biometrika*, 63, 581-592.

SEKAR, C.C., et DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

BIBLIOGRAPHIE

- ANDERSON, M. J., et FIENBERG, S. E. (1999). *Who Counts? The Politics of Census-Taking in Contemporary America*. New York: The Russell Sage Foundation.
- BELIN, T. (2001). Evaluation of unresolved enumeration status in 2000 Census Accuracy and Coverage Evaluation program. Rapport non publié, préparé par Datametrics, Inc., pour U.S. Census Bureau.
- BELIN, T., DIFFENDAL, G., MACK, S., RUBIN, D., SCHAFER, J., et ZASLAVSKY, A. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in reinterview. *Journal of the American Statistical Association*, 88, 1149-1166.
- CANTWELL, P. J. (2001). Accuracy and Coverage Evaluation Survey: Specifications for the missing data procedures. *DSSD Census 2000 Procedures and Operations Memorandum Series*, Q-62.
- CANTWELL, P. J., MCGRATH, D., NGUYEN, N., et ZELENAK, M. F. (2001). Accuracy and Coverage Evaluation: missing data results. *DSSD Census 2000 Procedures and Operations Memorandum Series*, B-7*.
- CHILBERS, D. (2000). The Design of the Census 2000 Accuracy and Coverage Evaluation. *DSSD Census 2000 Procedures and Operations Memorandum Series*, Chapter S-DT-1.
- FENSTERMAKER, D. (2000). The Accuracy And Coverage Evaluation: sample design summary. *DSSD Census 2000 Procedures and Operations Memorandum Series*, R-33.
- HAINES, D. (2003). A.C.E. Revision II results: changes in estimated net undercount. *DSSD A.C.E. Revision II Memorandum Series*, PP-58.
- HOGAN, H. (1993). The Post-Enumeration Survey: Operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H. (2003). L'évaluation de l'exactitude et de la couverture: théorie et conception. *Techniques d'enquête*, 29, 145-156.
- IKEDA, M., KEARNEY, A., et PETRONI, R. (1998). Missing data procedures in the Census 2000 Dress Rehearsal Integrated Coverage Measurement sample. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 468-473.
- KEARNEY, A., et IKEDA, M. (1999). Handling of missing data in the Census 2000 Dress Rehearsal Integrated coverage measurement sample. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 468-473.
- KEATHELY, D., KEARNEY, A., et BELT, W. (2001). ESCAP II, Analysis of missing data alternatives for the Accuracy and Policy II (ESCAP II) Rapport 12.
- KILLION, R. A. (2000). Measurement Error Reinterview Sample Selection. *Planning, Research, and Evaluation Division TXE/2010 Memorandum Series*, CM-MER-S-01.
- Les auteurs remercient Eric Schindler et Doug Olson d'avoir calculé les estimations selon le système dual et leurs erreurs-types sous les diverses méthodes de rechange, Tom Belin, de l'UCLA, d'avoir fourni les probabilités d'imputation sous les modèles de régression logistique et Mary Frances Zelelak et Ha Nguyen d'avoir produit des sommaires de la portée des données manquantes dans l'ACE. Le présent rapport est diffusé en vue de tenir les parties intéressées au courant des travaux de recherche et de favoriser la discussion. Les opinions exprimées sont celles des auteurs et ne représentent pas forcément celles du U.S. Census Bureau.
- ## REMERCIEMENTS
- Les auteurs remercient Eric Schindler et Doug Olson d'avoir calculé les estimations selon le système dual et leurs erreurs-types sous les diverses méthodes de rechange, Tom Belin, de l'UCLA, d'avoir fourni les probabilités d'imputation sous les modèles de régression logistique et Mary Frances Zelelak et Ha Nguyen d'avoir produit des sommaires de la portée des données manquantes dans l'ACE. Le présent rapport est diffusé en vue de tenir les parties intéressées au courant des travaux de recherche et de favoriser la discussion. Les opinions exprimées sont celles des auteurs et ne représentent pas forcément celles du U.S. Census Bureau.

diminution peut être attribuée presque exclusivement au domaine « Blancs non hispaniques ». Dans le cas de la méthode de rechange (6), fondée sur les données et les probabilités de la MBR, on observe une augmentation significative des estimations pour chaque domaine. Sous cette méthode, les différences relatives sont plus importantes que pour les méthodes de rechange précédentes, mais, en valeur absolue, sont toutes inférieures à 0,4 %. Plusieurs différences relatives sont supérieures à 0,3 % en valeur absolue : pour les Noirs non hispaniques, les Hispaniques et les Amérindiens hors réserve.

6. OBSERVATIONS

Les observations faites ici ont trait à la troisième catégorie de données manquantes nécessitant l'attribution de probabilités pour les cas non résolus de personnes dans l'ACE. Il est important de souligner que les méthodes de l'ACE ont été spécifiées bien avant l'exécution du recensement et de l'ACE. Les dates d'échec de la spécification des méthodes étaient précoces à cause (1) de l'échec d'un très grand nombre de la coordination des nombreuses activités distinctes, mais interdépendantes et (2) du besoin de mettre en place un processus transparent pour les décideurs, ainsi que les experts statistiques. Bien qu'on puisse en apprendre beaucoup au sujet des données manquantes et des structures de corrélation pertinentes en examinant les réponses telles qu'elles sont recueillies, la prise de décisions après avoir examiné les données pourrait avoir été considérée comme une manipulation des résultats d'une opération ayant des conséquences politiques graves.

Dans cette optique, on peut faire un retour en arrière et se rendre compte de divers moyens d'améliorer le processus – trop tard toutefois pour modifier les procédures. Cela ne signifie pas que nous n'avons pas réagi à l'information devenue disponible de façon imprévue durant le traitement des données. Nous savions que l'opération de suivi après l'appariement aiderait à résoudre certains cas, surtout ceux pour lesquels la situation de résidence réelle le jour du recensement était incertaine. De nombreux autres renseignements ont été recueillis durant ces interviews, mais nous ne nous attendions pas à en voir les détails. Cependant, grâce à des opérations intensives de saisie des données figurant sur les questionnaires de suivi au centre de traitement du Bureau, nous avons eu connaissance de certains renseignements supplémentaires durant l'opération de traitement des données manquantes. À ce moment-là, nous avons ajouté plusieurs groupes de code d'appariement non prévus au départ, à savoir le groupe 8 pour la situation de résidence, les groupes 11 et 12 pour la situation de recensement. Répartir les personnes entre ces groupes nous a permis d'attribuer des probabilités assez différentes – et, attribuées autrement.

gappe de la cellule comptant 24 personnes ayant un poids élevé incorrectement géocodé, tel que l'a décelé la MBS. La probabilité de recensement correct d'après la MBS pour le groupe de code d'appariement 11, « une ou plusieurs caractéristiques imputées », égale à 0,176, est un peu plus élevée que la valeur de 0,088 produite par l'ACE. La plupart des autres probabilités pour les situations de résidence, d'appariement et de recensement sont proches (écart de plus ou moins 0,03), comprises entre les valeurs de l'ACE et de la MBS; pour toutes les autres, l'écart est de plus ou moins 0,07.

Par contre, la MBR a été conçue pour évaluer l'erreur de collecte des données liée au processus d'appariement de l'ACE. Les personnes qui ont participé à la MBR ont été réinterviewées environ neuf mois après le jour du recensement pour recueillir des renseignements analogues à ceux recueillis durant les opérations de suivi de l'ACE, mais de façon plus détaillée. D'après la MBR, les probabilités d'être un résident ont tendance à être considérablement plus élevées pour les cellules du groupe de code d'appariement 8, mais plus faibles pour celles des groupes 3, 4 et 5 (représentant les non-appariements). La réduction a tendance à être importante dans les cellules où le groupe 8 a soutiré un plus grand nombre de cas aux groupes 3, 4 et 5. Soulignons que, dans le sous-groupe 3a, les cellules MBR sont assez petites. La cellule « Blancs non hispaniques, non-propriétaires » ne compte que 34 personnes non pondérées dont le cas n'est pas résolu, tandis que les trois autres cellules du groupe 3a comptent de 125 à 140 personnes non pondérées dont le cas est résolu. Les probabilités établies d'après la MBR pour la situation de recensement présentent un comportement comparable, celles pour les groupes 11 et 12 étant plus élevées et celles faibles. Pour la situation d'appariement, les probabilités obtenues d'après l'ACE et la MBR sont semblables, variant en général de 0,01 à 0,05.

Avant d'examiner les estimations selon le système dual sous les méthodes de rechange (5) (probabilités MBS) et (6) (probabilités MBR), il convient de souligner que, pour les comparaisons du tableau 9, seules les probabilités assignées aux cas non résolus ont été modifiées d'après les données recueillies lors de la MBS ou de la MBR. Bien que la situation évaluée de certaines personnes puisse avoir changé (par exemple, de non-appariement à appariement, ou de résident confirmé à résident non résolu) d'après les évaluations, leur situation n'a pas été modifiée lors du calcul de ces estimations, car l'objectif de l'exercice était d'examiner les diverses méthodes ou de déterminer comment elles influent sur la composition des méthodes de traitement des données manquantes des estimations selon le système dual.

Sous la méthode de rechange (5), fondée sur les données et les probabilités de la MBS, les estimations diminuent pour presque tous les domaines de population du tableau 9, quoique la baisse n'excède jamais 0,1 %. Cependant, cette

Tableau 9

Estimations selon le système dual pour diverses méthodes de rechange de traitement des données manquantes

Chaque cellule à la droite de la ligne verticale contient, par ordre, les estimations de a) la différence : estimation selon la méthode de rechange moins l'estimation de l'ACE, b) l'erreur-type de cette différence et c) la différence relative en pourcentage.

Différences estimées fondées sur six méthodes de rechange de traitement des données manquantes de l'ACE			
	(1)	(2)	(3)
Estimation de l'ACE (erreur-type)	Correction pour la non-interview avec cellules regroupées	d'imputation regroupées : échantillon P uniquement	d'imputation regroupées : échantillon F uniquement
	(1)	(2)	(3)
Blancs non hispaniques	-2 467	-32 324	-1 677
Blancs non hispaniques	(265 893)	(1 055)	(1 870)
	0,00 %	-0,02 %	0,00 %
Noirs non hispaniques	-3 495	-11 136	-119
Noirs non hispaniques	(34 210 774)	(1 290)	(1 328)
	-0,01 %	-0,03 %	0,00 %
Hispaniques	35 552 109	725	1 432
Hispaniques	(138 870)	(3 016)	(1 297)
	0,00 %	-0,02 %	-0,02 %
Autochtones hawaïens ou des îles du Pacifique	618 698	-98	88
Autochtones hawaïens ou des îles du Pacifique	(17 873)	(81)	(85)
	-0,02 %	-0,01 %	0,01 %
Asiatiques non hispaniques	10 056 009	709	-237
Asiatiques non hispaniques	(64 372)	(571)	(567)
	0,01 %	-0,03 %	-0,03 %
Amérindiens vivant dans les réserves	567 053	-245	61
Amérindiens vivant dans les réserves	(7 235)	(300)	(52)
	-0,04 %	-0,01 %	0,00 %
Amérindiens hors réserve	1 617 944	572	-96
Amérindiens hors réserve	(22 032)	(661)	(186)
	0,04 %	-0,02 %	-0,01 %
Domaines race-groupe ethnique			
Total - E.-U.	276 848 873	-4 299	-55 284
	(366 543)	(7 423)	(2 581)
	0,00 %	-0,02 %	0,00 %
Groupes d'âge			
Propriétaire	188 764 543	-2 237	-34 503
Propriétaire	(260 408)	(3 805)	(1 205)
	0,00 %	-0,02 %	0,00 %
Non-propriétaire	88 084 330	-2 063	-20 782
Non-propriétaire	(226 108)	(6 057)	(1 121)
	0,00 %	-0,02 %	0,00 %
Occupation du logement			
0 à 17 ans	73 076 071	2 924	-21 872
0 à 17 ans	(137 126)	(2 624)	(1 324)
	0,00 %	-0,03 %	0,00 %
18 à 49 ans	129 785 393	-2 721	-23 304
18 à 49 ans	(208 070)	(4 714)	(1 143)
	0,00 %	-0,02 %	0,00 %
50 ans et plus	73 987 409	-4 502	-10 108
50 ans et plus	(111 125)	(2 766)	(563)
	0,01 %	-0,01 %	0,00 %

font passer la valeur de 0,567 à 0,684. Comme le montre le tableau 9, l'effet sur les estimations selon le système dual est statistiquement significatif pour l'ensemble des Etats-Unis et pour presque toutes les ventilations présentes, sauf pour deux groupes race-groupe ethnique dont la taille est inférieure à un million de personnes. Par contre, les différences relatives ne semblent pas être très importantes, variant de 0,01 % à 0,04 %. Il n'est pas facile de déterminer quelle option de traitement des données manquantes produit les estimations s'approchant le plus des valeurs réelles inconnues.

Sous la méthode de rechange (3), les probabilités d'un recensement correct ont été recalculées en utilisant uniquement les groupes de code d'appareillement comme cellules d'imputation. Une variation importante des probabilités a été observée dans les cellules (originales) du groupe de code d'appareillement 3a. En ce qui concerne les estimations selon le système dual, des différences significatives n'ont été observées que pour deux des trois catégories d'âge et certains petits domaines race-groupe ethnique. À part pour ces domaines, toutes les différences en pourcentage sont inférieures à 0,01 %. Comme la méthode de rechange (4) s'appuie sur les probabilités recalculées à partir des échantillons P et E , ici, les estimations résultantes sont dominées par les résultats pour l'échantillon P et sont donc comparables à celles obtenues sous la méthode de rechange (2).

Les deux dernières méthodes de rechange emploient le même ensemble de cellules d'imputation que celles utilisées pour l'ACE, mais attribuent aux cas non résolus dans les échantillons P et E des probabilités éventuellement améliorées, telles que déterminées d'après l'ACE. Dans le cas de la méthode de rechange (5), les probabilités sont déterminées d'après l'erreur d'appareillement (MES pour Matching Error Study), tandis que dans le cas de la méthode de rechange (6), elles sont fondées sur la réinterview pour l'erreur de mesure (MER pour Measurement Error Reinterview). Chaque étude a été réalisée sur un ensemble de grappes d'évaluation constituant un sous-échantillon d'environ un cinquième de l'échantillon de grappes d'îlots de l'ACE. Pour des renseignements sur les plans de sondage de la MES et de la MER, consulter Petroni (2001) et Killian (2000).

L'objectif principal de la MES était d'évaluer les opérations d'appareillement de personnes de l'ACE. Des spécialistes de l'appareillement ont réparti les grappes d'évaluation et ont modifié les codes d'appareillement et de situation de personne finaux, au besoin. Aucune donnée supplémentaire n'a été recueillie pour la MES. Les probabilités dans les cellules d'imputation fondées sur la MES étaient, dans l'ensemble, semblables à celles attribuées lors de l'ACE, sauf pour la situation de résidence, dans le cas de la cellule du groupe de code d'appareillement 4. Blancs non hispaniques, non-proprétaires. Dans ce cas, la probabilité d'après la MES, 0,712, est nettement plus faible que la valeur de 0,911 produite par l'ACE. Ce résultat est dû à une

aucune différence entre les estimations présentées au tableau 9 pour cette méthode n'est statistiquement significative (supérieure à deux erreurs-types). Pareillement, excepté pour plusieurs domaines race-groupe ethnique comptant moins de deux millions de personnes, aucune différence relative n'est supérieure à 0,01 %.

Les méthodes de rechange (2), (3) et (4) ont été élaborées après avoir examiné les effets des variables utilisées dans les cellules d'imputation sur les probabilités assignées résultantes. D'après les probabilités assignées des tableaux 4 et 6, il est évident que le groupe de code de résidence et de recensement, par opposition à la situation de recensement. Pourtant, il semble que la subdivision des cellules d'après les variables démographiques, comme « Blancs non hispaniques » par opposition à « Autre », ne permet pas une aussi bonne différenciation. Pour étudier l'effet des variables démographiques sur l'imputation, on a assigné de nouvelles probabilités aux cas non résolus sans les utiliser. Plus précisément, on a combiné tous les cas résolus et non résolus sur l'ensemble des cellules pour les catégories Blancs non hispaniques et Autre (situation de résidence et situation de recensement), pour les groupes de code d'appareillement 3a et 3b (situation de résidence et de recensement), et pour « Aucune caractéristique imputée » et « Au moins une caractéristique imputée » (situation d'appareillement et de recensement); les variables dérivées d'après les opérations de l'ACE, c'est-à-dire le groupe de code d'appareillement, le code d'appareillement d'adresse d'unité de logement et la situation de déménagement ont été retenues. La méthode de rechange (2) consiste à appliquer la méthode de correction uniquement à l'ensemble plus petit de cellules dans l'échantillon P , autrement dit uniquement aux cas non résolus de situation de résidence et de situation d'appareillement; la méthode de rechange (3) consiste à appliquer la méthode de correction uniquement à l'échantillon E (situation de recensement) et la méthode de rechange (4) consiste à

appliquer aux deux échantillons. Sous la méthode de rechange (2), la variation la plus importante des probabilités d'être un résident attribuées aux cas non résolus s'observe dans les quatre cellules d'imputation (originales) du groupe 3a, ne concernant que 96 personnes dont la situation n'était pas résolue. Dans la plupart des autres cellules pour la situation de résidence (plus de 99 % des cas), les probabilités ne varient que très faiblement. Pour les probabilités d'appareillement, on n'observe une différence importante que pour la cellule « personne n'ayant pas de déménagement, unité non apparée ou imputée » contenant 421 cas non résolus. La variable différenciant le nombre de cas imputés semble avoir eu un effet ici; si l'on regroupe ses deux sous-cellules « imputées », la probabilité assignée à la cellule « une ou nettement plus grand de cas résolus sans imputation, qui

différences entre les estimations causées par les diverses méthodes de traitement des données manquantes.

5.1 Résultats d'une évaluation précoce

Durant les mois qui ont suivi la diffusion des estimations provisoires selon le système dual d'après l'ACE, d'autres méthodes de traitement des données manquantes ont été étudiées. Les raisons de cette étude étaient multiples, à savoir estimer la variation qui pourrait résulter de la variation dans l'analyse de l'erreur totale et de la fonction de perte pour les estimations selon le système dual de l'ACE et l'étude de la viabilité de méthodes de traitement des cas de données manquantes non ignorables pour présentes dans Keahley, Keaney et Bell (2001), nous nous limiterons à les résumer ici.

Trois méthodes comportant la correction pour la non-interview ont été examinées. Dans la première, les cellules ont été définies différemment pour la correction, par ajout de variables comme la race, l'origine hispanique, le mode d'occupation du logement et la taille du ménage, telles que déterminées d'après un appartenance aux enregistrements du fichier de recensement. Cette méthode a généralement produit des estimations selon le système dual plus grandes. Deux autres méthodes de correction pour la non-interview n'ont eu aucun effet apparent sur les estimations. Dans l'une, c'est-à-dire une correction pour la non-interview par la méthode du plus proche voisin, on a ajouté le poids d'un ménage non interviewé à celui du ménage interviewé le plus proche dans le fichier trié. Dans l'autre, on a qualifié de « tardives » les 30 % d'interviews de l'ACE réalisées en dernier lieu. Le poids des unités non interviewées a été ajouté uniquement au poids des interviews tardives. Ces méthodes visaient à tirer parti de l'hétérogénéité attendue des unités résultant de la proximité géographique ou du moment de la réponse à l'ACE.

Les autres méthodes décrites dans Keahley et coll. (2001) traitent les cas non résolus de situation de résidence, d'appartenance ou de recensement. Une approche fondée sur les données « tardives » s'appuie sur des renseignements recueillis uniquement auprès des 30 % d'interviews réalisées en dernier lieu dans l'échantillon *P* ou d'unités de logement nécessitant un suivi pour la non-réponse dans l'échantillon *E*. En soi, cette approche ne semblait pas influencer sur les estimations selon le système dual. Les autres méthodes étudiées comportent des modèles de régression logistique pour prédire les probabilités pour les cas non résolus. Premièrement, on a appliqué aux cas non résolus de situation de résidence, d'appartenance et de dénombrement un modèle logistique ignorable, celui décrit plus haut (section 4.3) dans Bellin (2001), qui a eu tendance à produire des estimations selon le système dual plus faibles (chiffre inférieur de 47 à 481 pour l'ensemble des États-Unis). Cependant, il semble que les probabilités plus faibles (en moyenne) de recensement correct attribuées aux

5.2 Analyses d'autres méthodes de rechange

À la présente section, nous présentons les différences entre les estimations selon le système dual pour six méthodes de rechange, décrites et justifiées plus loin. Les résultats sont présentés au tableau 9 pour l'ensemble des États-Unis et pour la ventilation selon la race-groupe ethnique, le mode d'occupation du logement et l'âge. Pour une définition précise des domaines de race-groupe ethnique, voir Kostasich (2001). Il convient de souligner qu'une petite partie de la population des États-Unis ne faisait pas partie de l'univers de l'ACE. Pour chaque méthode, les trois chiffres donnés sont a) la différence entre l'estimation par la méthode de rechange et l'estimation de l'ACE, b) l'erreur-type de cette différence et c) la différence relative en pourcentage.

Dans le cas de la méthode de rechange (1), on considère de nouveau la méthode de correction pour la non-interview appliquée à l'ACE pour corriger les cellules contenant un nombre assez faible d'interviews complètes (voir la section 3). Dans le cas de cette méthode de rechange, au lieu de répartir les poids des unités non interviewées sur une gamme étendue de cellules, on a regroupé les cellules contenant un trop petit nombre d'interviews avec les cellules voisines et on a de nouveau calculé les facteurs de correction pour la non-interview pour les cellules nouvellement créées. Sauf pour les Noirs non hispaniques,

Tableau 8
Probabilités moyennes de recensement correct pour l'imputation par cellule et la régression logistique

Situation de recensement		Echantillon E		Groupe de code d'appariement		Nombre de cas non résolus		Probabilité moyenne attribuée		Régression logistique	
Imputation		par cellule		résolus		cas non résolus		Régression		logistique	
1. Appariements nécessitant un suivi		711		0,977		0,986		0,967		0,967	
2. Appariements éventuels		305		0,968		0,963		0,974		0,974	
3. Non-appariements partiels du ménage		2 191		0,962		0,963		0,973		0,973	
4. Non-appariements complets du ménage où l'unité de logement a été appariée; ménages compatibles		4 813		0,967		0,926		0,917		0,917	
5. Non-appariements de ménages incompatibles; unités de logement non incluses dans le suivi pour la non-réponse		532		0,973		0,961		0,954		0,954	
6. Non-appariements de ménages incompatibles; unités de logement incluses dans le suivi pour la non-réponse		779		0,917		0,982		0,999		0,999	
7. Non-appariements complets du ménage, où l'unité de logement n'a pas été appariée		381		0,954		0,982		0,999		0,999	
8. Cas résolus avant le suivi		179		0,99		0,982		0,999		0,999	
9. Information insuffisante pour l'appariement		0		---		---		---		---	
10. Cas de recherche étendue cible		2 902		0,918		0,679		0,679		0,679	
11. Personnes éventuellement fictives		1 690		0,064		0,077		0,077		0,077	
12. Personnes déclarées comme vivant ailleurs le jour du recensement		3 132		0,225		0,28		0,28		0,28	

cellule ont été attribuées conformément aux équations (3) et (4), qui intègrent le poids ACE, si bien que les cas éliminés de l'échantillon de l'ACE n'ont pas été inclus dans les calculs. Le modèle de régression logistique a été exécuté sur les données non pondérées et incluait les 8 298 cas du groupe 10, ce qui fait baisser la probabilité d'un recensement correct prédite pour les 2 902 personnes pour laquelle la situation de recensement n'était pas résolue.

5. EFFET DE CERTAINES AUTRES MÉTHODES DE TRAITEMENT DES DONNÉES MANQUANTES SUR LES ESTIMATIONS SELON LE SYSTÈME DUAL

À la section précédente, les probabilités prévues ont été comparées pour deux méthodes de traitement des cas non résolus. Cependant, en dernière analyse, l'effet de méthodes concurrentes s'observe dans les estimations selon le système dual résultantes. À la présente section, nous comparons plusieurs autres méthodes que celles utilisées dans le cadre de l'ACE pour traiter les données manquantes en nous fondant sur les estimations résultantes. Lorsque celles-ci diffèrent considérablement, le choix de la meilleure méthode n'est pas évident. Il convient de souligner que les estimations de l'ACE diffusées par le U.S. Census Bureau en mars de 2001 ont été révisées à la suite d'analyses plus approfondies (Haines 2003). Quoique les données de l'ACE soient entachées d'erreurs et que les estimations de l'ACE ne devraient généralement pas être utilisées, on estime qu'elles conviennent pour évaluer les

où leur probabilité d'appariement sous régression logistique aurait été beaucoup plus élevée. Donc, pour ce petit ensemble de 134 cas, la variable logistique, groupe de code d'appariement, prend une valeur incorrecte et le modèle prédit une probabilité – beaucoup trop faible – fondée sur un grand nombre de cas résolus dans les groupes 3, 4 ou 5 qui étaient en réalité des non-appariements, mais avaient été sélectionnés pour le suivi principalement pour résoudre leur situation de résidence et non leur situation d'appariement. Les probabilités prévues d'appariement sont aussi fort différentes pour le groupe 8. Cependant, vu qu'il n'y a que six cas non résolus, l'effet sur l'estimation devrait être minime.

Lorsqu'on compare les probabilités moyennes de recensement correct selon le groupe de code d'appariement au tableau 8, on ne constate presque aucune différence, sauf pour le groupe 10, contenant les cas de recherche étendue cible. Pour ce groupe, la probabilité moyenne attribuée par l'imputation par cellule, 0,918, est beaucoup plus élevée que celle prédite par le modèle de régression logistique, soit 0,679. La pondération permet d'expliquer cette différence. Dans l'échantillon E, parmi les 32 334 personnes admissibles à l'opération de recherche étendue cible, 8 298 (toutes comprises dans le groupe de code d'appariement 10) ont été écartées de l'échantillon pour réduire les coûts et se sont vues attribuer un poids de l'ACE nul. Durant l'opération d'appariement, il n'a pas été essayé de déterminer si les 8 298 cas avaient été recensés correctement ou non, ceux-ci ayant simplement été gardés dans le fichier de données comme des enregistrements de recensement incorrects. Les probabilités fondées sur l'imputation par

4.3 Comparaison des probabilités sous imputation par cellule et sous régression logistique

La comparaison des probabilités attribuées aux cas non résolus selon des procédures différentes peut être intéressante. Belin (2001) donne une comparaison de ce genre pour un modèle de régression logistique tenant compte de 186 prédicteurs pour les situations de résidence et d'appartement et de 202 prédicteurs pour la situation de recensement. Les variables englobaient la plupart de celles utilisées pour l'imputation par cellule décrites à la section 4.2, ainsi que les caractéristiques démographiques individuelles, comme l'âge, le sexe et la relation avec la personne de référence, l'information au sujet de l'interview de l'ACE, comme le fait que la réponse ait été obtenue par procuration, les caractéristiques régionales locales, comme celles de région urbaine ou non urbaine, et les interactions entre les variables. Puisqu'on a ajusté les modèles aux cas résolus sélectionnés pour le suivi, puis qu'on les a appliqués aux cas non résolus pour prédire une probabilité, on peut ne pas tenir compte de l'effet du modèle, en ce sens que la situation non résolue n'est pas considérée comme une covariable dans le modèle sous-jacent. (Voir Rubin 1976.)

Les tableaux 7 et 8 résument les probabilités moyennes attribuées aux cas non résolus sous la méthode d'imputation par cellule de l'ACE et la modélisation logistique calculées sur les divers groupes de code d'appartement. Rappelons qu'on a calculé les probabilités par la méthode d'imputation par cellule à partir de données pondérées comme dans (3), mais qu'on a exécuté les modèles de régression logistique sur des données non pondérées. On a calculé la moyenne non pondérée des probabilités prévues selon les deux méthodes sur l'ensemble des personnes représentant un cas non résolu. À une exception près que nous mentionnerons plus tard, les probabilités et les estimations produites par l'ACE étaient généralement semblables qu'on utilise des données pondérées ou non pondérées, car l'échantillonnage avait été planifié de façon à éviter une gamme étendue de poids.

La comparaison des méthodes ne révèle presque aucune différence entre les probabilités moyennes attribuées pour une situation de résidence. Ce résultat n'est pas étonnant, puisque l'imputation par cellule s'appuie sur le groupe de code d'appartement (entre autres variables) pour la définition des cellules. Le tableau est différent pour la situation d'appartement. Rappelons que le groupe de code d'appartement n'a pas été utilisé pour l'imputation par cellule car presque tous les cas non résolus d'appartement (98,3 % de 7 826; 7 506 avant l'opération de suivi et 1 87 de plus après le suivi) ne fournissaient pas suffisamment de renseignements pour procéder à l'appartement. Les probabilités attribuées pour les deux premiers groupes varient légèrement selon la méthode utilisée. Par contre, pour les groupes 3, 4 et 5, qui englobent tous les cas de non-appartement avant le suivi, les probabilités moyennes sont plus élevées sous imputation par cellule (0,893, 0,770 et 0,616) et très faibles sous régression logistique (0,050, 0,010 et 0,070). Parmi les 156 cas compris dans les trois cellules, 134 correspondaient à des personnes auxquelles on avait donné un code initial indiquant un « non-appartement »; plus tard, il a été établi correctement que les renseignements concernant la personne étaient insuffisants pour procéder à l'appartement. Dans presque chaque cas, l'intervieweur de l'ACE avait enregistré un nom comme « Enfant Jones », « José Ne sait pas » ou « Inconnu Smith ». Les cas de ce genre auraient dû être repérés par un commis avant l'appartement et recevoir un code initial d'information insuffisante. Au lieu de cela, une tentative d'appartement aux enregistrements du recensement a eu lieu et a échoué. Si cette erreur n'avait pas été commise, les personnes concernées auraient été placées dans le groupe 7,

Tableau 7
Probabilités moyennes d'être un résident et d'appartement pour l'imputation par cellule et la régression logistique

Groupe de code d'appartement	Échantillon P		Situation de résidence		Situation d'appartement	
	Probabilité moyenne attribuée	Nombre de cas non résolus	Probabilité moyenne attribuée	Nombre de cas non résolus	Probabilité moyenne attribuée	Nombre de cas non résolus
1. Appartements nécessitant un suivi	0,941	767	0,989	0,983	0,848	4
2. Appartements éventuels	0,837	352	0,970	0,962	0,889	131
3. Non-appartements partiels du ménage	0,050	1 306	0,956	0,951	0,893	71
4. Non-appartements complets du ménage	0,010	1 610	0,917	0,926	0,770	36
5. Non-appartements, ménage incompatible	0,070	1 455	0,940	0,927	0,616	49
6. Cas résolus avant le suivi	0,940	129	0,990	0,990	0,842	23
7. Information insuffisante	0,880	7 506	0,844	0,851	0,835	7 506
8. Personnes fictives, vivant ailleurs	0,041	2 402	0,148	0,167	0,655	6

Tableau 6
Cellules d'imputation pour la résolution de la situation de recensement dans l'échantillon E

Groupe de code d'appartement de l'échantillon E		Aucune caractéristique	Au moins une caractéristique
		imputée ¹	imputée
1.	Appartements nécessitant un suivi	0,977	0,977
2.	Appartements éventuels	0,968	0,968
3a.	Non-appartements partiels du ménage; groupe des 18 à 29 ans, enfant de la personne de référence	0,871	0,908
3b.	Non-appartements partiels du ménage; autres que dans 3a	0,974	0,96
4.	Non-appartements complets du ménage où l'unité de logement est apparée; ménages compatibles	Blancs non hispaniques 0,965	Autre 0,958
5.	Non-appartements de ménages incompatibles; pour les unités de logement non comprises dans le suivi régulier pour la non-réponse	0,975	0,974
6.	Non-appartements de ménages incompatibles; unités de logement comprises dans le suivi régulier de la non-réponse	0,914	0,926
7.	Non-appartements complets du ménage, où l'unité de logement n'a pas été apparée lors de l'appariement des unités de logement	Blancs non hispaniques 0,959	Autre 0,947
8.	Cas résolus avant le suivi	Blancs non hispaniques 0,995	Autre 0,990
9.	Renseignements insuffisants pour l'appariement		0 (attribué par définition)
10.	Cas de recherche étendue ciblée ²	0,928	0,858
11.	Personnes éventuellement fictives	0,058	0,088
12.	Personnes déclarées comme vivant ailleurs le jour du recensement	0,229	0,210

¹ Parmi les caractéristiques suivantes : âge, sexe, mode d'occupation du logement, race ou origine hispanique. ² Par recherche étendue ciblée, on entend une opération sur le terrain réalisée pour réduire la variance des estimations selon le système dual causée par les erreurs de géocodage des grappes. Pour plus de renseignements, voir Navarro et Olson (2001).

cas de l'échantillon P, les personnes considérées comme éventuellement fictives ou ayant été déclaré comme vivant ailleurs le jour du recensement durant les opérations de suivi ont été placées les premières dans les groupes 11 et 12, respectivement. Puis, on a placé les autres personnes faisant partie de l'échantillon E dans le groupe de code d'appariement approprié, tel que défini dans le tableau. On a subdivisé le groupe 3 en deux sous-groupes, comme pour la détermination de la situation de résidence pour l'échantillon P. Autrement dit, toute personne de 18 à 29 ans qui était l'enfant de la personne de référence a été classée séparément. Les autres caractéristiques utilisées pour définir les cellules sont la présence ou l'absence de caractéristiques imputées, tel que défini dans les cellules d'imputation pour la situation d'appariement et le fait que la personne était Blanche non hispanique ou appartenait à toute autre combinaison race-groupe ethnique. Il convient de souligner que, conformément aux procédures de l'ACE, on a automatiquement attribué une probabilité de recensement correcte nulle à toute personne faisant partie de l'échantillon E pour laquelle on ne disposait pas de renseignements suffisants pour procéder à l'appariement (groupe 9).

Les enregistrements de personnes pour lesquels la valeur avait été imputée pour au moins une variable démographique (âge, sexe, mode d'occupation du logement, race ou origine hispanique) ont été regroupés lors de la résolution des cas pour la situation d'appariement. Selon des études non publiées, du moins parmi les cas résolus lors de la répétition générale du recensement, la présence d'imputation de ces caractéristiques est associée négativement à la propension d'être un appartement. On ne s'est donc pas fondé sur ces variables pour séparer les personnes étant sorties d'une unité correspondant à un non-appartement ou à un ménage incompatible pour s'assurer d'obtenir un nombre raisonnable de cas résolus dans chaque cellule utilisée pour estimer la proportion d'appariement. Dans l'échantillon E, les cas non résolus de situation de recensement ont été traités tel qu'exposé plus haut. Voir le tableau 6.

Comme pour la situation de résidence pour les personnes comprises dans l'échantillon P, un déterminant clé de la situation de recensement est le groupe de code d'appariement des personnes comprises dans l'échantillon E, même si les groupes de code d'appariement ont été définis différemment pour les deux échantillons. Comme dans le

différent de cellules, comme le montre le tableau 5. Les cas confirmés de non-résidents ont été exclus des calculs des probabilités d'appariement.

Lors de la répétition générale du recensement, pour déterminer la probabilité pour les cas d'appariement non résolus, on n'a utilisé qu'une seule cellule d'imputation dans chaque empiètement géographique. L'analyse subséquente (Kearney et Ikeda 1999) a montré que la situation de déménagement (personne n'ayant pas déménagé c. personne sortante) permettait de bien faire la distinction entre les appartements et les non-appartements parmi les cas résolus. Donc, pour l'ACE de 2000, on a utilisé la situation de déménagement pour définir les cellules d'imputation pour la situation d'appariement. Le code d'appariement de l'adresse de l'unité de logement renvoie à l'appariement initial entre les unités de logement figurant sur la liste établie indépendamment (ACE) et la liste d'adresses du recensement; les unités de logement incompatibles, repérées durant les activités d'appariement des personnes de l'ACE, sont celles pour lesquelles les listes de membres du ménage établies pour le recensement et pour l'ACE donnaient deux listes entièrement différentes.

Il convient de souligner que 98,3 % des cas non résolus d'appariement (7 693 sur 7 826) correspondaient à des personnes pour lesquelles on ne disposait pas de suffisamment de renseignements pour procéder à l'appariement. Comme nous l'avons mentionné plus haut, la plupart de ces entregistrements ne contenaient même pas un nom valide et presque tous (7 506) n'ont pas été sélectionnés pour le suivi. En outre, leur taux de caractéristiques pour lesquelles des données manquaient était nettement supérieur à la moyenne. Par conséquent, ces enregistrements ne fournissaient que peu d'information prédictive lors de la création des cellules d'imputation pour la détermination de la situation d'appariement. On a évité d'utiliser des variables comme l'âge et le groupe ethnique, pour lesquelles la probabilité d'imputation était plus forte et dont la qualité risquait d'être douteuse.

référence. Nombre de ces jeunes personnes fréquentaient un collège, partageaient une résidence avec des collègues, ou emménageaient dans la résidence de leurs parents ou la quittaient. La classification et l'analyse par arbre de régression appliquées aux données de la répétition générale du recensement réalisée en 1998 donnaient à penser que cette combinaison de caractéristiques permettrait de bien faire la distinction en ce qui a trait à la situation de résidence. Le groupe 3b englobait toutes les autres personnes du groupe 3.

On a calculé la probabilité d'être un résident pour les cas non résolus de l'échantillon *P* selon la méthode décrite plus haut, sauf pour les personnes classées dans le groupe de code d'appariement 7, c'est-à-dire les personnes pour lesquelles on ne disposait pas de renseignements suffisants pour procéder à l'appariement. Sur cette ligne du tableau 4, il n'y avait essentiellement aucun cas résolu duquel on pouvait extraire une probabilité d'avoir été un résident le jour du recensement. À cause du manque d'information – on ne disposait même pas d'un nom valide pour la plupart de ces cas – ces personnes n'ont pas été traitées durant l'opération d'appariement et n'ont pas été retenues pour le suivi. Pour déterminer la probabilité d'avoir été un résident pour ces cas, on a calculé une proportion pondérée de résidents le jour du recensement (valeurs 1 et 0) parmi les cas résolus des groupes de code d'appariement 1 à 5 et 8, séparément pour chacune des quatre catégories Mode d'occupation du logement x race-groupe ethnique. Puis, on a exclu de ce calcul les personnes dont le cas avait été résolu avant le suivi (groupe 6). Les observations faites lors de la répétition générale indiquaient que, du point de vue des caractéristiques démographiques et opérationnelles, les personnes du groupe 7 avaient tendance à ressentir davantage à celles des groupes 1 à 5 et 8 qu'à celles du groupe 6.

La question des appartements non résolus a été traitée comme celle des situations de résidence non résolues dans (3) et (4), en remplaçant la situation de résidence par la situation d'appariement, mais en utilisant un ensemble

Tableau 5
Cellules d'imputation pour la résolution de la situation d'appariement dans l'échantillon *P*

Code d'appariement de l'adresse de l'unité de logement		Unité de logement appariée		Unité de logement non appariée ou ménage incompatible	
Situation de déménagement	Personne n'ayant pas déménagé	Personne n'ayant pas déménagé	Personne sortante	Personne n'ayant pas déménagé	Personne sortante
Aucune caractéristique	Aucune caractéristique	Aucune caractéristique	Aucune caractéristique	Aucune caractéristique	Aucune caractéristique
0,945 imputée	0,901 imputée	0,901 imputée	0,791 imputée	0,567 caractéristique imputée	0,516

¹ Parmi les caractéristiques suivantes : âge, sexe, mode d'occupation du logement, race ou origine hispanique.

Tableau 4
Cellules d'imputation pour la résolution de la situation de résidence dans l'échantillon P

Échantillon P		Groupe de code d'appartenance		Blancs non hispaniques		Propriétaires		Blancs non hispaniques		Non-propriétaires	
1.	Appartements nécessitant un suivi	0,982	0,973	0,986	0,993	0,991	0,972	0,972	0,966	0,983	0,928
2.	Appartements possibles										
3a.	Non-appartements partiels de ménages nécessitant un suivi; groupe des 18 à 29 ans, enfant de la personne de référence	0,755	0,971	0,901	0,959	0,969	0,928	0,928	0,883	0,959	0,928
3b.	Non-appartements partiels de ménages nécessitant un suivi; autres que dans 3a	0,956	0,971	0,971	0,959	0,969	0,928	0,928	0,883	0,959	0,928
4.	Non-appartements complets du ménage nécessitant un suivi; ménages comparables	0,920	0,943	0,943	0,911	0,914	0,928	0,928	0,883	0,959	0,928
5.	Non-appartements de ménages incompatibles	0,910	0,927	0,927	0,945	0,954	0,928	0,928	0,883	0,959	0,928
6.	Cas résolus avant le suivi	0,993	0,990	0,990	0,990	0,988	0,928	0,928	0,883	0,959	0,928
7.	Renseignements insuffisants pour l'appartenance (moyenne pondérée de colonne pour les groupes 1 à 5 et 8)	0,813	0,867	0,867	0,844	0,872	0,928	0,928	0,883	0,959	0,928
8.	Personnes éventuellement fictives ou déclarées comme vivant ailleurs le jour du recensement	0,119	0,123	0,123	0,177	0,157	0,928	0,928	0,883	0,959	0,928

$$(4) \quad I_{res,j} = \begin{cases} 1 & \text{si } j \in R(i) \\ P(res)_i, & \text{autrement} \end{cases}$$

Pour calculer les nombres estimés de personnes n'ayant pas déménagé et de personnes sortantes dans l'échantillon P qui figurent dans l'équation (2), N_{nm} et N_{om} , respectivement, on applique le poids de la personne et l'indicateur $I_{res,j}$ à toutes les personnes n'ayant pas déménagé et toutes les personnes sortantes, respectivement, dans toutes les cellules. On procède de façon analogue pour déterminer le nombre d'appartements, M_{nm} ou M_{om} , et donc, p_{match} de même que p_{cc} , dans le cas de la situation de recensement. Au moment de la répétition générale du recensement en 1998, la méthode d'imputation par cellule pour calculer la probabilité d'être un résident pour les cas non résolus n'a été appliquée qu'à trois cellules, à savoir les personnes retenues pour un suivi, les personnes ne nécessitant pas de suivi et les personnes pour lesquelles les renseignements étaient insuffisants pour procéder à l'appartenance. Pour la troisième cellule, qui ne contenait aucun cas résolu, on a attribué une proportion fondée sur tous les cas résolus dans les deux premières cellules. Les résultats de la répétition générale (Kearney et Ikeda 1999) donnaient à penser que la subdivision de l'échantillon P pour former les divers groupes de code d'appartenance serait utile. D'autres études et discussions laissent entendre qu'il fallait ajouter d'autres variables démographiques dans les groupes de code d'appartenance. La plus grande taille de l'échantillon de l'ACB réalisée en 2000 a permis d'utiliser un ensemble plus grand de cellules d'imputation.

Pour l'ACB de 2000, les groupes de code d'appartenance 1 à 7 ont été définis d'après les codes d'appartenance et d'autres variables dérivées avant les opérations de suivi, tel que décrit dans Childers (2000). Le groupe 8 a été formé de l'opération de suivi en temps voulu pour les procédures de traitement des données manquantes de l'ACB. (Aux termes du calendrier original, cette information aurait été disponible trop tard pour pouvoir être utilisée.) Après l'opération de suivi, le code de personne éventuellement fictive ou déclarée comme résidant ailleurs le jour du recensement a été attribué à un petit nombre de personnes comprises dans l'échantillon P. Parmi les cas résolus compris dans ce groupe, la probabilité d'être un résident était nettement plus faible que pour les cas résolus compris dans les autres groupes. Donc, les personnes satisfaisant aux conditions pour être classées dans le groupe 8 y ont été placées pour commencer, puis chacune des personnes restantes a été placée comme il convenait dans l'un des sept premiers groupes.

On a utilisé deux catégories de mode d'occupation du logement, à savoir propriétaire et non-propriétaire. Les personnes ont également été classées dans l'une de deux catégories de race-groupe ethnique, à savoir Blancs non hispaniques et autres. Les personnes appartenant à plusieurs races (par exemple une personne indiquant qu'elle était Blanche et Asiatique) ont été classées dans le second groupe. Le groupe de code d'appartenance 3, cas de non-appartenance partiel du ménage, a été subdivisé en deux sous-groupes. Le premier, 3a, comprenait les personnes du groupe 3 de 18 à 29 ans inscrites sur la liste de membres du ménage de l'ACB comme étant l'enfant de la personne de

4.2 Attribution de probabilités aux cas non résolus

Dans l'ACE, on a utilisé une forme d'imputation par cellule pour attribuer une probabilité aux cas échantillonnés pour lesquels la situation de résidence, d'appartement ou de recensement n'était pas résolue. Toutes les personnes comprises dans l'échantillon – cas résolus et non résolus – ont été placées dans des groupes appelés cellules d'imputation en fonction de caractéristiques opérationnelles et démographiques. Pour chaque type de situation, on a utilisé des variables différentes pour définir les cellules. Dans chaque cellule d'imputation, la moyenne pondérée des valeurs 1 et 0 (représentant, par exemple, un appartement et un non-appartement, respectivement) a été calculée pour les cas résolus, puis imputée à tous les cas non résolus figurant dans la cellule. Des renseignements supplémentaires sont donnés plus loin.

Lors de l'enquête postcensitaire de 1990, on a utilisé un modèle de régression logistique hiérarchique pour calculer les probabilités d'appartement et de recensement correct pour les cas pour lesquels des données manquaient. (Étant donné la méthode utilisée en 1990 pour traiter les personnes ayant déménagé, la situation de résidence jouait un rôle différent à ce moment-là.) Le modèle et certains résultats sont exposés dans Belin et coll. (1993).

Durant les années 1990, le Censur Bureau prévoyait produire en 2000 des estimations de recensement corrigées pour chacun des 50 États (et le District de Columbia) au moyen de données recueillies uniquement dans l'État en question. Cette approche avait deux types de répercussions sur la stratégie de traitement des cas non résolus. Premièrement, dans chaque État, on aurait disposé d'un nombre moins grand de données, c'est-à-dire de cas résolus, pour construire un modèle de régression logistique. Deuxièmement, il aurait fallu examiner et vérifier 153 modèles différents, c'est-à-dire des modèles distincts pour les situations de résidence, d'appartement et de recensement dans chaque État. Comme le calendrier de production de l'ACE n'accordait que trois semaines environ pour aborder tous les aspects des données manquantes, on a estimé que l'adoption d'une méthode de traitement des cas non résolus plus facile à mettre en œuvre et à vérifier réduirait le risque de ne pas produire les estimations selon le système dual dans les délais imposés. L'imputation par cellule offrait la simplicité souhaitée, mais son exactitude comparativement à l'imputation par régression logistique a été évaluée lors d'essais subséquents.

Durant les essais du recensement réalisés en 1995 et en 1996, on a traité certains types de situation non résolue par régression logistique et d'autres par la méthode d'imputation par cellule. Cette dernière méthode avait été utilisée uniquement lors de la répétition générale du recensement en 1998 (Ikeda, Kearney et Petroni 1998), à l'époque où le Censur Bureau se préparait encore à produire des estimations indépendantes pour chaque État. Les données de ces essais ont indiqué que la méthode exacte de calcul des probabilités pour les cas non résolus (appartenance, résident

ou recensement correct) n'avait qu'un effet très faible sur les estimations selon le système dual. Pour des renseignements détaillés sur cette étude, consulter Petroni (1997, 1998a, 1998b et 1998c).

Étant donné le jugement rendu par la Cour suprême des États-Unis en 1999 (*Dept. of Commerce v. U.S. House of Representatives*), le Censur Bureau a modifié la conception de l'enquête et éliminé la contrainte voulant que les estimations corrigées soient fondées uniquement sur les données recueillies dans chaque État. Cependant, des réserves persistaient quant à l'application d'une méthode de régression logistique n'ayant pas été testée lors de la répétition générale. En outre, rien ne garantissait que les logiciels disponibles permettraient d'exécuter adéquatement les modèles de régression logistique sur les ensembles de données de la taille de l'échantillon complet de l'ACE (de 640 000 à 750 000 personnes). Compte tenu de ces réserves et des résultats de l'étude sur l'exactitude relative, le Censur Bureau a décidé d'utiliser la procédure plus simple d'imputation par cellule pour résoudre les cas pour lesquels la situation manquait dans l'ACE.

Pour illustrer la façon dont l'imputation par cellule a été appliquée à l'ACE, on peut examiner la situation de résidence, la méthode a été appliquée de façon analogue pour la situation d'appartement et pour la situation de recensement. Premièrement, on a placé toutes les personnes n'ayant pas déménagé et les personnes sortantes comprises dans l'échantillon P dans un certain nombre de cellules d'imputation d'après certaines caractéristiques opérationnelles et démographiques, tel que précisé au tableau 4; les personnes entrantes ont été écartées, puisque leur probabilité d'être un résident le jour du recensement était nulle par définition. Parmi les cas résolus dans la cellule i , représentée par l'ensemble $R(i)$, on a défini une variable indicatrice de la situation de résidence comme étant égale à 1 si la personne j faisait partie du ménage le jour du recensement et 0, autrement. Puis, dans la cellule i , on a calculé la proportion pondérée de résidents le jour du recensement, soit :

$$(3) \quad P(res)_i = \frac{\sum_{j \in R(i)} w_j \cdot 1_{res,j}}{\sum_{j \in R(i)} w_j} = \frac{\sum_{j \in R(i)} w_j \cdot f_{res}(i)}{\sum_{j \in R(i)} w_j}$$

où w_j est le poids de la personne j intégrant toutes les étapes d'échantillonnage. Ensuite, on a attribué la proportion $P(res)_i$ à chaque personne dont le cas n'était pas résolu dans la cellule i , c'est-à-dire chacune des 15 082 personnes (2,3 % de 653 337) dont la situation de résidence n'était pas résolue. (La seule exception a eu lieu pour le groupe de code d'appartement 7, comme nous l'expliquons plus loin.) Le tableau 4 donne les probabilités d'être un résident attribuées à l'intérieur des cellules. Cette attribution définit, pour tous les cas, résolus et non résolus, un indicateur « étendu » pouvant prendre toutes les valeurs comprises entre 0 et 1 :

4.1 Cas non résolus et leur fréquence

L'une des composantes de l'estimateur à système dual donné par l'équation (1) est \hat{p}_{match} , c'est-à-dire la proportion estimée de l'échantillon P qui correspond à des personnes recensées. Dans (2), pour \hat{p}_{match} , lors de l'estimation du nombre de personnes (N_{nm} , N_{om}) ou d'appariements (M_{nm} , M_{om}) parmi les personnes n'ayant pas de déménagement et les personnes sortantes, on a considéré uniquement les résidents, le jour du recensement, des grappes d'îlots échantillonnées; par exemple, les personnes résidant habituellement dans une maison de soins infirmiers ont été omises du calcul. Donc, pour chaque personne comprise dans l'échantillon P , il a fallu déterminer la situation de résidence et la situation d'appariement. Quand les opérations de suivi ont été achevées, toutes les personnes comprises dans l'échantillon P remplissant les critères d'admissibilité pour l'appariement aux enregistrements du recensement ont été réparties en trois catégories selon leur situation de résidence le jour du recensement dans l'îlot échantillonné, c'est-à-dire résident, non-résident et cas non résolus, c'est-à-dire cas pour lesquels les renseignements recueillis ne suffisaient pas pour déterminer la situation de résidence. En outre, pour chaque résident confirmé ou éventuel (cas non résolu) le jour du recensement compris dans l'échantillon P , on a déterminé s'il s'agissait d'un appartement, d'un non-appariement ou d'un

Le tableau 3 donne les fréquences pour la situation de résidence et la situation d'appariement dans l'échantillon P et pour la situation de dénombrement dans l'échantillon E . Il donne aussi les résultats pour les personnes n'ayant pas de déménagement et les personnes sortantes dans l'échantillon P . L'examen du tableau montre que la proportion de cas non résolus est assez faible, soit 2,3 % pour la situation de résidence, 1,2 % pour la situation d'appariement et 3,0 % pour la situation de recensement. (Les taux pondérés sont 2,2 %, 1,2 % et 2,6 %, respectivement.) Lors de l'enquête postcensitaire de 1990, le taux de cas d'appariement non résolus était de 1,9 % et le taux de cas de recensement non résolus était de 2,4 %. (La situation de résidence n'avait pas été définie de façon comparable en 2000.) Il faut néanmoins interpréter les résultats avec prudence, car les définitions de plusieurs situations étaient légèrement différentes en 1990.

Tableau 3

Fréquences des situations finales pour les échantillons P et E (non pondérés)

Échantillon P		Nombre total de personnes ¹		Résident		Non-résident		Résident non résolu		Taux de résidents pour les cas résolus	
Situation finale de résidence		Personnes ²		Appariement		Non-appariement		Appariement non résolu		Taux d'appariements pour les cas résolus	
Situation de dénombrement	Personne n'ayant pas de déménagement	627 992	96,6 %	1,7 %	1,7 %	1,7 %	1,7 %	1,7 %	1,7 %	98,3 %	98,3 %
	Personne sortante	25 345	75,2 %	7,5 %	7,5 %	7,5 %	7,5 %	7,5 %	7,5 %	91,0 %	91,0 %
	Total - E.-U.	653 337	95,8 %	1,9 %	1,9 %	1,9 %	1,9 %	1,9 %	1,9 %	98,1 %	98,1 %
Échantillon E		Nombre total de personnes		Recensement correct		Recensement incorrect		Recensement non résolu		Taux de recensements corrects pour les cas résolus	
Situation de dénombrement		Personnes		Recensement		Recensement		Recensement non résolu		Taux de recensements corrects pour les cas résolus	
Situation de dénombrement	Personne n'ayant pas de déménagement	617 490	91,1 %	8,0 %	8,0 %	8,0 %	8,0 %	8,0 %	8,0 %	91,9 %	91,9 %
	Personne sortante	23 455	67,8 %	21,7 %	21,7 %	21,7 %	21,7 %	21,7 %	21,7 %	75,8 %	75,8 %
	Total - E.-U.	704 602	92,6 %	4,4 %	4,4 %	4,4 %	4,4 %	4,4 %	4,4 %	95,5 %	95,5 %

¹ Personnes comprises dans l'échantillon P admissibles pour l'appariement aux enregistrements de recensement.

² Résidents confirmés ou éventuels dans l'échantillon P .

Les résultats des opérations d'interview de l'ACE sont présentés au tableau 2. Parmi les 261 969 unités de logement occupées le jour du recensement, 7 794 (3,0 %) représentent des non-interviews. Les nombres correspondants pour le jour de l'interview de l'ACE sont 267 155 et 3 052 (1,1 %).

Tableau 2
Situation d'interview des ménages lors de l'ACE

Situation d'interview des ménages lors de l'ACE (non pondérés)		Taux de non-interview ¹		Taux de non-interview = Non-interviews / (interviews + non-interviews)	
jour du recensement	Nbre %	interviews	Nbre %	interviews	Nbre %
Nbre total d'unités de logement	300 913	254 175	84,5	264 103	87,8
Unités occupées	28 472	7 794	2,6	3 052	1,0
Unités supprimées	10 472	3,5	4 096	1,4	

Comme la situation d'interview d'une unité de logement pouvait être différente le jour du recensement et le jour de l'interview de l'ACE, différentes corrections pour la non-interview ont été nécessaires pour chaque jour de référence. En général les deux corrections consistaient à répartir les poids des unités non interviewées sur les unités interviewées dans la même cellule de non-interview, c'est-à-dire le croisement de la grappe d'îlots échantillonnée et du type d'adresse de base, défini comme étant une unité unifamiliale, une unité multifamiliale (comme les immeubles à appartements et les condominiums) ou toutes les autres unités. On aurait pu utiliser d'autres caractéristiques, connues pour toutes les unités de logement, pour définir les cellules. Cependant, on a décidé de tirer parti de l'homogénéité locale typique, et du fait que les personnes qui vivent en appartement, par exemple, ont de nombreuses caractéristiques communes, comme la taille du ménage ou la propension à déménager, qui sont associées aux probabilités de capture lors du recensement.

On a utilisé la correction pour la non-interview fondée sur la situation des unités de logement le jour du recensement pour corriger la pondération des personnes n'ayant pas déménagé et des personnes ayant quitté les unités de logement (personnes sortantes). De la même façon, on a utilisé la correction pour la non-interview le jour de l'interview de l'ACE pour corriger la pondération des personnes venues s'installer dans les unités de logement (personnes entrantes). À l'intérieur d'une cellule de non-interview, on a calculé le facteur de correction pour le

Quand le nombre non pondéré d'unités non interviewées dans une cellule de non-interview partielle était égal à plus du double du nombre non pondéré d'unités interviewées, on a réparti les poids des unités non interviewées dans cette cellule sur les unités interviewées dans un ensemble plus grand de cellules de non-interview. Ce remède n'a été nécessaire que pour 65 cellules pour la correction pour le jour du recensement et pour 13 cellules pour la correction pour le jour de l'interview de l'ACE. La procédure prescrite diffère du regroupement habituel des cellules peu peuplées, mais permet de traiter simplement ce genre de cellules de façon automatisée. La capacité de procéder ainsi était très importante dans des conditions où les délais étaient courts et où il était impossible de prédire quelles cellules contiendraient un nombre trop faible d'interviews. Aux fins d'évaluation, on a recalculé ultérieurement les poids des unités de logement selon un processus de regroupement des cellules et on les a comparés aux poids déterminés par la méthode de l'ACE. De nouveau, étant donné les taux faibles de non-interview, la pondération était la même pour la plupart des unités et approchante pour les autres. L'effet sur les estimations selon le système dual résultantes est présenté à la section 5.2.

4. ATTRIBUTION DE PROBABILITÉS AUX CAS NON RÉSOLUS

Quand toutes les activités de suivi de l'ACE ont été achevées, les renseignements nécessaires pour calculer les composantes de l'estimateur à système dual donné par l'équation (1) étaient encore insuffisants pour une petite fraction de l'échantillon de l'ACE. Ces cas ont été dits « non résolus ».

Representatives, 525 U.S. 316, 1999), on a estimé que modifier de nouveau la procédure visant les personnes ayant déménagé à une étape aussi avancée avant le recensement introduirait des risques inacceptables. Étant donné la procédure susmentionnée pour tenir compte des personnes ayant déménagé, il existe deux situations d'interview pour chaque unité de logement, l'une fondée sur la situation de l'unité de logement le jour du de l'ACE. Les unités inoccupées ou supprimées de la liste des unités de logement admissibles (par exemple, parce qu'elles avaient été démolies ou utilisées uniquement à des fins commerciales), et certaines unités situées à des emplacements spéciaux n'ont été considérées ni comme une interview ni comme une non-interview. Le tableau 1 donne un exemple fictif de grappe d'îlots. Il montre comment la situation d'une unité de logement le jour du recensement et le jour de l'interview de l'ACE a été déterminée.

La méthode de dénombrement des personnes ayant déménagé utilisée pour l'ACE différerait de celle utilisée pour l'enquête postcensitaire de 1990. Cette année-là, on a utilisé les personnes entrantes pour estimer le nombre de personnes ayant déménagé et leur taux d'appariement. Pour ce dernier, il a fallu réappairer les personnes entrantes à leur adresse le jour du recensement. Cette procédure a été modifiée pour les essais de recensement réalisés au cours des années 1990 pour tenir compte de l'utilisation prévue de l'échantillonnage pour les non-répondants au recensement. Quand la Cour suprême des États-Unis a rendu un jugement contre le plan d'échantillonnage en 1999 (*Department of Commerce v. United States House of*

Tableau 1
Exemple de correction pour la non-interview

Considérons une grappe d'îlots contenant neuf unités de logement, ayant toutes le même type d'adresse de base, par exemple, maison unifamiliale, tel que décrit ci-dessous			
Unité de logement	Poids	Situation réelle	Situation (et information provenant) de l'interview de l'ACE
Situation le jour du recensement			
Situation le jour de l'interview de l'ACE			
1	100	Résident le 1 ^{er} avril 2000 et au moment de l'interview de l'ACE	Interviewée lors de l'ACE
2	100	Résident le 1 ^{er} avril et au moment de l'interview de l'ACE	Voisin interviewé (procuration) lors de l'ACE
3	100	Résident le 1 ^{er} avril et au moment de l'interview de l'ACE	Aucune personne interviewée lors de l'ACE
4	100	Inoccupé le 1 ^{er} avril, résident au moment de l'interview de l'ACE	Interviewée lors de l'ACE, situation Inoccupée
5	100	Inoccupé le 1 ^{er} avril, résident au moment de l'interview de l'ACE	Interviewée lors de l'ACE, situation Non-interview
6	100	Inoccupé le 1 ^{er} avril, résident au moment de l'interview de l'ACE	Aucune personne interviewée lors de l'ACE
7	100	Résident le 1 ^{er} avril, inoccupé au moment de l'interview de l'ACE	Information obtenue par procuration
8	100	Résident le 1 ^{er} avril, inoccupé au moment de l'interview de l'ACE	Aucune information sur la situation
9	100	Résident le 1 ^{er} avril, résident au moment de l'interview de l'ACE	Interviewée lors de l'ACE, situation Interview

Dans cette cellule de non-interview (grappe d'îlots échantillonnée x type d'adresse de base), on aurait appliqué les corrections pour la non-interview suivantes aux personnes occupant les unités de logement interviewées :

- (1) à la pondération des personnes n'ayant pas déménagé et des personnes sortantes, la correction pour la non-interview le jour du recensement = $800 / 400 = 2,0$
- (2) à la pondération des personnes entrantes, la correction pour la non-interview le jour de l'interview de l'ACE = $700 / 500 = 1,4$.

3. CORRECTION POUR LA NON-INTERVIEW

La correction pour la non-interview n'a été faite que pour l'échantillon P , dans le cas du recensement (et, donc, de l'échantillon E), diverses procédures ont été utilisées pour tenir compte des personnes vivant dans toutes les unités de logement connues. Le petit nombre d'unités de logement pour lesquelles l'information a été recueillie par procuration, souvent auprès d'un voisin ou du gestionnaire de l'immeuble, ont été considérées comme des interviews non-interview. Comme des personnes sont venues s'installer dans les unités de logement ou en ont déménagé le jour du recensement et le moment de l'interview de l'ACE, le Census Bureau a dû tenir compte de la situation de déménagement – personne sortante (out-mover), de déménagement (non-mover) ou personne n'ayant pas déménagé (in-mover) – pour toutes les personnes faisant partie de l'échantillon P , ainsi que de la situation le jour de l'interview pour les deux interviews différentes. Les personnes sortantes sont celles qui vivaient dans l'unité de logement en question le jour du recensement, mais l'avaient quittée avant le jour de l'interview de l'ACE. La situation est inverse pour les personnes entrantes. Les personnes n'ayant pas déménagé sont celles qui vivaient dans l'unité lors de chaque interview. Au moment de l'interview de l'ACE, en une seule interview, on a posé des questions pour déterminer qui vivait dans le ménage au moment de l'interview de l'ACE et qui y vivait le jour du recensement. On a attribué une situation de déménagement à chaque personne comprise dans l'échantillon et on a créé deux listes de membres pour chaque ménage, à savoir la liste de membres le jour du recensement et la liste de membres le jour de l'interview de l'ACE.

Dans le cas de l'ACE, on a utilisé les personnes entrantes pour estimer le nombre de personnes ayant déménagé dans l'échantillon P , et les personnes sortantes pour estimer le total pondéré de l'échantillon P , autrement dit le dénominateur de p_{match} dans l'équation (2), est égal à la somme pondérée de toutes les personnes n'ayant pas déménagé et des personnes entrantes. Par ailleurs, on estime le nombre pondéré d'appariements dans l'échantillon P en ajoutant le nombre d'appariements parmi les personnes n'ayant pas déménagé au produit du nombre de personnes entrantes et du taux d'appariement pour les personnes sortantes, soit :

$$(2) \quad \hat{p}_{\text{match}} = \frac{M_{nm} + N_{nm} \times \frac{N_{om}}{M_{om}}}{M_{nm} + N_{nm} + N_{om}}$$

où N (personnes) et M (appareils) présentent les indices nm , im et om , qui représentent les personnes n'ayant pas déménagé (non-movers), les personnes entrantes (in-movers) et les personnes sortantes (out-movers), respectivement.

contenues dans l'échantillon de grappes d'îlots sélectionné pour l'ACE. Après le recensement – mais en n'utilisant aucune information recueillie lors du recensement – le Census Bureau a interviewé indépendamment les personnes comprises dans l'échantillon de l'ACE et obtenu une liste de personnes vivant dans les unités le jour du recensement, c'est-à-dire le 1^{er} avril 2000. Puis, ces résultats ont été appariés (comparés) aux chiffres de recensement dans ces grappes d'îlots pour estimer le nombre de personnes n'ayant pas été recensées. Dans les grappes d'îlots échantillonnées, les unités dénombrées indépendamment dans le cadre de l'ACE ont été définies comme formant l'échantillon P et celles dénombrées lors du recensement, comme formant l'échantillon E .

Sur le même échantillon de grappes d'îlots, on a procédé à des comparaisons et à des analyses pour estimer la proportion d'enregistrements de recensement corrects, c'est-à-dire d'enregistrements complets, uniques et effectués à l'emplacemement approprié. Les dénombrements incorrects incluent les personnes comptées en double ou fictives, ou celles qui n'auraient pas dû être dénombrées à l'adresse en question, par exemple, parce que leur résidence habituelle était située ailleurs, comme un dortoir de collège. L'estimateur à système dual résultant est

$$(1) \quad \hat{N} = (C - I) \hat{p}^{ce} \left(\frac{1}{\hat{p}_{\text{match}}} \right),$$

Les estimations selon le système dual ont été calculées séparément pour certains sous-groupes de population appelés strates à posteriori. Puis, on a utilisé les estimations à posteriori pour déterminer les facteurs de correction à appliquer à toutes les personnes dénombrées au recensement d'après leur strate à posteriori particulière. Enfin, on a calculé les dénombrements corrigés pour toute région géographique par totalisation des dénombrements corrigés sur l'ensemble des strates à posteriori contenues dans la région. Pour des renseignements plus détaillés sur les opérations sur le terrain de l'ACE et l'estimation selon le système dual en général, consulter Childers (2000) et Hogan (1993, 2003), respectivement.

(par exemple, conjoint(e), enfant) ou si les renseignements sur le conjoint ou la conjointe de la personne sont disponibles. Par conséquent, on a utilisé les distributions nationales des donneurs conditionnellement aux covariables pertinentes pour imputer des valeurs pour l'âge et le sexe. Comme les caractéristiques de l'imputation pour l'ACE sont semblables à celles de l'imputation lors de l'enquête postcensitaire qui a suivi le Recensement de 1990, nous ne discutons pas davantage des méthodes et des résultats dans le présent article.

Le troisième type de données manquantes correspond aussi à des données manquantes pour une question. Pour un petit nombre de participants à l'ACE, un nombre insuffisant de renseignements ont été recueillis pour déterminer la situation de résidence (si oui ou non, le jour du recensement, la personne résidait dans la grappe d'échantillon(e) ou la situation d'appartenance (si oui ou non la personne correspondait effectivement à une personne recensée). De la même façon, pour certaines personnes recensées, on ne possédait pas assez d'information pour déterminer si elles l'avaient été correctement. Les cas de ce genre sont dits « non résolus ». Cependant, cette information est nécessaire pour calculer les estimations selon le système dual. Pour résoudre de tels cas, on a attribué une probabilité d'être un résident (ou d'appartenance) moyenne calculée d'après un ensemble de cas résolus présentant des caractéristiques similaires.

Certains de ces méthodes, qui sont décrites de façon plus détaillée plus loin, ont été appliquées sous la même forme durant l'enquête postcensitaire de 1990 et durant les tests effectués au cours des années 1990. La principale exception est l'attribution d'une probabilité aux cas non résolus de situation de résidence, d'appartenance ou de dénombrement. Dans le cas de l'enquête postcensitaire et, parfois, pour des évaluations particulières au cours des années 1990, on a calculé ces probabilités au moyen d'un modèle de régression logistique. La méthode appliquée pour l'ACE de 2000 se fondait sur moins d'information que certaines méthodes de recharge, comme la régression logistique, mais on l'avait choisie parce qu'elle était plus simple à appliquer et à vérifier, étant donné les échecs sévères de l'ACE.

Le taux de données manquantes observé pour l'ACE était relativement faible, ce qui a réduit le risque d'erreur supplémentaire dans les estimations.

- Les taux de non-interview des ménages étaient de 3,0 % et 1,1 % (non pondérés), respectivement, le jour du recensement et le jour de l'interview de l'enquête.
- Les taux d'imputation pour les cinq caractéristiques de l'ACE requises pour l'estimation selon le système dual variaient de 1,4 % à 2,5 % (non pondérés et pondérés).

lors de l'estimation.

À la section qui suit, nous présentons de façon générale les fréquences et les proportions non pondérées. Sauf indication contraire, les nombres pondérés sont très proches. Toutefois, les probabilités attribuées aux cas non résolus aux tableaux 4, 5 et 6 sont les poids réels utilisés

2. BRÈVE DESCRIPTION DE L'ENQUÊTE ET DE L'ESTIMATION SELON LE SYSTÈME DUAL

Grâce à l'évaluation de l'exactitude et de la couverture (ACE), le Census Bureau visait à mesurer et à corriger le sous-dénombrement net historique observé pour le recensement des États-Unis (Anderson et Fienberg 1999, page 29). Comme le Recensement de 2000, l'ACE couvrirait l'entière de la population. (Un échantillon distinct a été sélectionné et analysé pour Puerto Rico.) Aux fins de l'enquête, on a sélectionné un échantillon d'environ 300 000 unités de logement dans 11 303 grappes d'îlots (Fenstermaker 2000; Childers 2000).

L'estimation de la couverture de la population grâce à l'ACE est fondée sur l'estimation selon le système dual, méthode qui s'inspire de la méthodologie de capture-recapture (Peterson 1896; Sekar et Deming 1949). Supposons qu'on ne considère que les unités de logement

Traitement des données manquantes dans l'enquête d'évaluation de l'exactitude et de la couverture de 2000

PATRICK J. CANTWELL et MICHAEL IKEDA¹

RÉSUMÉ

L'enquête d'évaluation de l'exactitude et de la couverture a été réalisée pour estimer la couverture du Recensement des États-Unis de 2000. Après l'achèvement des opérations sur le terrain, il a fallu prendre des mesures pour traiter plusieurs types de données manquantes en vue d'appliquer l'estimateur à système dual. Certaines unités de logement n'avaient pas été interviewées. Le cas échéant, on a conçu deux méthodes de correction pour la non-interview d'après un même ensemble d'interviews, une pour chaque point dans le temps. En outre, il a fallu déterminer la situation de résidence, d'appartenance ou de recensement de certains répondants. Les méthodes appliquées par le passé ont été remplacées pour pouvoir respecter les délais plus courts pour calculer et vérifier les estimations. Le présent article décrit la portée des données manquantes dans l'enquête et les méthodes de traitement appliquées, compare ces dernières à d'autres méthodes passées et courantes, et donne un résumé analytique des procédures, y compris la comparaison des estimations démographiques selon le système dual sous d'autres méthodes de traitement des données manquantes. Comme les niveaux de données manquantes étaient faibles, il semble que l'utilisation des autres méthodes n'aurait pas affecté considérablement les résultats. Cependant, on constate certains changements dans les estimations.

MOTS CLÉS : Imputation par cellule; correction pour la non-interview; régression logistique; estimation selon le système dual.

1. INTRODUCTION

Après le Recensement des États-Unis de 2000, le Censur Bureau a réalisé l'enquête d'évaluation de l'exactitude et de la couverture (ACE pour *Accuracy and Coverage Evaluation*). L'enquête avait deux objectifs, à savoir 1) déterminer le sous-dénombrement net à l'échelle nationale et pour divers domaines démographiques et géographiques grâce à une méthode statistique appelée estimation selon le système dual et 2) produire des chiffres de population révisés pouvant être utilisés pour faire la correction pour ce sous-dénombrement net – si les chiffres révisés étaient jugés plus exacts que les dénombrements initiaux au recensement (Hogan 2003).

Durant le processus d'interview et de suivi des répondants de l'ACE, certains ménages ont été manqués et certains renseignements nécessaires pour le calcul des estimations selon le système dual n'ont pas été recueillis auprès d'autres répondants échantillonnés. Le présent article décrit le niveau des données manquantes et les méthodes utilisées dans le cadre de l'ACE pour résoudre le problème et donner certains résultats et évaluations. Il convient de souligner que l'expression « données manquantes » s'entend des données manquantes après l'exécution de toutes les tentatives de suivi sur le terrain. Ces activités incluent les efforts multiples en vue de procéder aux interviews, l'utilisation de commis et de techniciens ayant reçu une formation poussée pour résoudre les cas, et le suivi des cas pour lesquels une deuxième interview pouvait fournir les renseignements supplémentaires requis.

L'ACE a donné lieu à trois catégories de données manquantes. Premièrement, certains ménages n'ont pas été interviewés parce qu'on n'a pu prendre contact avec eux ou qu'ils ont refusé de participer à l'interview. L'élément qui rend la situation différente dans le cas de l'ACE est que, pour chaque unité de logement échantillonnée, on a appliqué deux corrections pour la non-interview, l'une pour la non-interview le jour du recensement, et l'autre, pour la non-interview le jour de l'interview de l'ACE. Comme nous le montrerons, la nécessité de faire deux corrections reflète la façon différente de traiter les personnes qui sont venues s'installer dans un endroit (personnes entrantes, ou *in-movers*) et les personnes ayant quitté cet endroit (personnes sortantes, ou *out-movers*) dans l'estimation selon le système dual.

Le deuxième type de données manquantes est survenu quand l'information sur un ménage ou une personne était disponible, mais que les données sur les caractéristiques démographiques particulières nécessaires pour l'estimation selon le système dual n'avaient pas été recueillies. Pour les données manquantes sur le mode d'occupation du logement (propriétaire c. non-propriétaire), la race et l'origine hispanique, on a recouru à une forme d'imputation hot-deck par le plus proche voisin pour profiter des corrélations souvent observées entre les personnes vivant géographiquement à proximité l'une de l'autre. En général, les valeurs pour l'âge et le sexe sont géographiquement moins regroupées, mais peuvent souvent être projetées d'après des caractéristiques particulières, comme la relation de la personne avec la personne de référence du ménage

¹ Patrick J. Cantwell et Michael Ikeda, statisticiens mathématiciens, U.S. Census Bureau, Statistical Research Division, Washington, D.C. 20233-9100.

- MULE, T. (2001). ESCAP II: Person Duplication in Census 2000. Executive Steering Committee for A.C.E. Policy II, Rapport 20.
- MULE, T. (2002). Further Study of Person Duplication Statistical Matching and Modeling Methodology. DSSD A.C.E. Revision II Memorandum Series PP-51.
- NASH, F.F. (2001). ESCAP II: Analysis of Census Imputations. Executive Steering Committee for A.C.E. Policy II, Rapport 21.
- PETERSEN, C.G.J. (1896). The Yearly Immigration of Young People into the Limfjord from the German Sea. Rapport de Danish Biological Station. 6, 1-48.
- RAGLIN, D. (2002). ESCAP II: Effect of Excluding Reinstated Census People from the A.C.E. Person Process. Rapport 13, <http://www.census.gov/dmd/www/pdf/Report13.PDF>
- ROBINSON, J.G., AHMED, B., DAS GUPTA, P. et WOODROW, K. (1993). Estimates of population coverage in the 1990 united states census based on demographic analysis. *Journal of the American Statistical Association*. 88, 1061-77.
- ROBINSON, J.G. (2001). ESCAP II: Demographic Analysis Results. Executive Steering Committee for A.C.E. Policy II, Report 1.
- SCHENKER, N. (1988). Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986. *Techniques d'enquête*. 14, 93-104.
- SEKAR, C.C., et DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*. 44, 101-115.
- U.S. CENSUS BUREAU (1985). Evaluating Census of Population and Housing, Statistical Training Document, ISP-TR-5, Washington, D.C.
- U. S. CENSUS BUREAU (2000). Statement on the Feasibility of Using Statistical Methods to Improve the Accuracy of Census 2000.
- U. S. CENSUS BUREAU (2003). Technical Assessment of A.C.E. Revision II, March 12, 2003, <http://www.census.gov/dmd/www/pdf/ACETechAssess.pdf>
- WETROGAN S.I., et CRESCE A.R. (2001). ESCAP II: Characteristics of Census Imputations. Executive Steering Committee for A.C.E. Policy II, Rapport 22.
- WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*. 81, 338-346.

BIBLIOGRAPHIE

- BELIN, T.R. (1993). Évaluation des sources de variation dans le couplage d'enregistrements au moyen d'une expérience factorielle. *Techniques d'enquête*, 19, 15-33.
- CANTWELL, P., et IKEDA, M. (2003). Traitement des données manquantes dans l'enquête d'évaluation de l'exactitude. *Techniques d'enquête*, 29, 2, sous presse.
- CHILBERS, D. (2001). Accuracy and Coverage Evaluation: The Design Document. DSSD Census 2000 Procedures and Operations Memorandum Series, Chapitre S-D-1 (Révisé). ESCAP I (2001). Report of the Executive Steering Committee for Accuracy and coverage Evaluation Policy. 1 Mars 2001. (Voir www.census.gov/dmd/www/pdf/Esca2.pdf)
- ESCAP II (2001). Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy on Adjustment for Non-Redistributing Uses. 17 Octobre 2001. (Voir www.census.gov/dmd/www/pdf/Recomend2.pdf)
- ESCAP II (2001). Census Person Duplication and the Corresponding A.C.E. Enumeration Status. Executive Steering Committee for A.C.E. Policy II, Rapport 6.
- FAY, R. (2002). Probabilistic models for detecting census person duplication. *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- FELDPAUSCH, R. (2001). ESCAP II: Census Person Duplication and the Corresponding A.C.E. Enumeration Status. Executive Steering Committee for A.C.E. Policy II, rapport 6.
- GONZALEZ, M. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section*, American Statistical Association, 73, 7-15.
- GONZALEZ, M., et HOZA, C. (1978). Small-area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 361, 7-15.
- GRIFFIN, R. (2000). Accuracy and Coverage Evaluation Survey: Dual System Estimation. DSSD Census 2000 Procedures and Operations Memorandum Series Q-20.
- HAINES, D. (2001). Accuracy and Coverage Evaluation Survey: Computer Specifications for Person Synthetic Estimation (U.S.). Re-issue of Q-30. DSSD Census 2000 Procedures and Operations Memorandum Series Q-46.
- HOGAN, H. (1992). The 1990 Post-Enumeration Survey: An overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The Post-Enumeration Survey: Operations and Results. *Journal of American Statistical Association*, 88, 423.
- HOGAN, H. (2001). Accuracy and Coverage Evaluation Survey: Effect of Excluding Late Census Adds. DSSD Census 2000 Procedures and Operations Memorandum Series Q-43. <http://www.census.gov/dmd/www/pdf/Q-43.pdf>
- MARKS, E.S., SELTZER, W. et KROTKI, K.J. (1974). *Population Growth Estimation*. New York : Population Council.
- MARKS, E.S. (1979). The Role of Dual System Estimation in Census Evaluation. In Recent Developments in PGE, (K. Krotki). University of Alberta Press. 156-188.
- Le présent article présente les résultats de travaux de recherche et d'analyse entrepris par les employés du Census Bureau. Les opinions émises sont celles de l'auteur et ne représentent pas nécessairement celles du Census Bureau.

REMERCIEMENTS

Leur enfant vivait à la maison même si les instructions du recensement indiquent clairement de ne pas le faire. Dans les cas de « garde conjointe », les deux parents peuvent systématiquement déclarer que l'enfant vit dans chacun des deux ménages. Les voisins déclareront, sans aucun doute, que quelqu'un « vit ici » alors que la personne est au collège, à l'armée, en prison ou dans une résidence secondaire. Ces erreurs de déclaration surviennent malgré les nombreuses questions détaillées et spécifiques d'approfondissement au sujet du lieu de résidence habituel posées par les intervieweurs de l'ACE.

La recherche approfondie d'enregistrements de recensement en double décrite plus haut constitue la preuve principale du résultat erroné de l'ACE. Cependant, d'autres preuves ont été recueillies, y compris une étude fondée sur détail dans le document intitulé « Executive Steering Committee on A.C.E. Policy » (ESCAP) du Census Bureau (voir ESCAP I 2001, ESCAP II 2001).

Selon ces évaluations, 4,7 millions d'enregistrements de recensement incorrects n'ont pu être décelés lors de l'ACE (U.S. Census Bureau 2003, page iv). En outre, l'étude a probablement donné lieu à l'identification incorrecte du lieu de résidence d'un grand nombre de personnes dans l'échantillon P , ce qui a donné des appariements ainsi que des non-appariements erronés. Un programme à grande échelle d'analyse et d'estimation mis en place par le Census Bureau a produit le surdénombrement estimatif de 1,3 million mentionné plus haut. Cependant, ce programme était adapté uniquement aux circonstances spéciales du réappariement, de la réinterview et de la recherche des enregistrements en double réalisés dans le cadre de l'enquête postcensitaire de 2000. Le lecteur qui souhaite plus de renseignements est invité à consulter U.S. Census Bureau (2003).

Le présent article décrit la théorie du DSE et explique comment cette théorie a été mise en application dans le cadre de l'EPIC en général, et de l'ACE en particulier. Il décrit aussi les approximations nécessaires lors des applications réelles et les types d'erreurs qui peuvent survenir.

Il précise avec quelle minute chaque approximation doit être contrôlée. De toute évidence, l'ACE n'a pas permis de dénombrer convenablement le grand nombre d'enregistrements en double au Recensement de 2000. La cause principale de cet échec a été l'impossibilité d'obtenir des déclarations exactes du lieu habituel de résidence même en posant de nombreuses questions d'approfondissement. Cependant, la théorie et le plan de sondage exposés ici devraient être utiles lors de la mise en place de tout programme de mesure de la couverture dans l'avenir.

Jusqu'à présent, nous avons accepté la classification dans les strates j , comme étant fixe. En pratique, certaines personnes sont classées dans des strates différentes au recensement et lors de l'enquête. Par exemple, l'âge déclaré d'une femme pourrait être 28 ans au moment du recensement et 31 ans au moment de l'enquête, ce qui la placerait dans des strates à posteriori

différentes.

partie de l'échantillon. Nous savons que ce cela s'est produit parce qu'après le recensement et l'ACE, nous avons pu procéder à une recherche et à un appariement à l'échelon national. Cet exercice nous a permis de rechercher les enregistrements de recensement en double, même lorsque les éléments de la paire étaient à des milles de distance. La recherche a été possible parce que, pour la première fois, presque tous les noms vus par le recensement ont été saisis électroniquement. (Voir Fay 2002; Mule 2001, 2002.) Nous avons ainsi pu constater, par exemple, que nombre de personnes classifiées d'après l'échantillon *E* de l'ACE comme étant « recensées correctement » avaient également été recensées ailleurs, y compris dans un autre ménage ou dans un logement collectif.

Ce genre d'erreur de déclaration n'est habituellement pas grave pour l'appariement. Le nom, l'adresse, le mois et le jour de la naissance, la composition du ménage et les liens entre les membres du ménage sont des données beaucoup plus importantes que l'âge, la race, voire même parfois le sexe. Donc, dans l'exemple susmentionné, en supposant qu'il y ait un appariement, nous aurions une femme de 28 ans recensée correctement dans l'échantillon *E* et une femme de 31 ans recensée correctement dans l'échantillon *P*. Nous voyons que l'erreur de classification a deux effets. Dans la mesure où les probabilités réelles de sous-échantonnement sont homogènes par rapport aux caractéristiques réelles, l'erreur de classification introduit une hétérogénéité (et un biais dû à l'hétérogénéité) dans les cellules d'estimation observées. Il en est ainsi même si les déclarations faites au moment du recensement et au moment de l'enquête concordent, car l'erreur de classification peut introduire de sous-groupes imbriqués dans les strates à posteriori où les probabilités d'inclusion dans chaque

La non-concordance des déclarations au recensement et au moment de l'enquête pose un problème pour l'estimateur synchrone, ainsi que pour le DSE. On peut le constater facilement en ne tenant pas compte des imputations lors du recensement ni des enrégistrement de recensement incorrects. Dans ces conditions, le facteur de correction de la couverture est égal à l'inverse du taux d'appariement (N_{11}^t/N_{12}^t) , où t représente la strate a posteriori. Si la classification dans la strate a posteriori n'est pas la même alors le recensement que pour l'enquête, nous appliquons pour le taux (estime pour un groupe) à un groupe quelconque peu différent. Si l'erreur de classification peut être ignorée au niveau de la strate a posteriori, elle pourrait être importante localement. Lors de l'ACE, en guise de mesure du possible, de définir les strates a posteriori, d'après des variables ayant une forte variabilité de déclaration.

L'ACE ET CONCLUSION

Alors qu'elle semblait bien conçue et bien exécutée, l'ACE n'a même pas permis de mesurer approximativement l'erreur de couverture au recensement des Etats-Unis de 2000. La raison principale semble avoir été la non-validité

5. ESTIMATION SYNTHÉTIQUE

5.1 Le modèle synthétique et le modèle à système dual

Jusqu'à présent, nous avons examiné le DSE réel. Cependant, comme nous l'avons mentionné à la section 2, nous utilisons un estimateur synthétique pour répartir le sous-dénombrement net mesuré entre les zones locales et les petits groupes.

Dans le cas de l'ACB, la ventilation a été fondée sur les mêmes variables de stratification à posteriori que le DSE proprement dit. L'estimation synthétique est basée sur les hypothèses que (1) le DSE estime la population réelle et (2) à l'intérieur des strates à posteriori, la population réelle est répartie proportionnellement au chiffre de recensement avant ajustement (prévu).

Manifestement, à un certain niveau, la deuxième hypothèse ne peut être vérifiée qu'en ce qui concerne les chiffres de recensement prévus. Autrement dit, même si, à l'intérieur des strates à posteriori, toutes les personnes avaient une probabilité identique d'être comptées lors du recensement, nous observerions des résultats différents selon l'ilot. Le DSE sous-jacent modélise explicitement le sous-dénombrement sous forme de processus stochastique. À mesure que les secteurs deviennent plus grands, deux choses se produisent. Premièrement, l'effet stochastique, ou l'« effet d'ilot » aléatoire, commence à s'uniformiser autour d'une valeur moyenne. Deuxièmement, l'effet du sous-dénombrement réel provenant d'une série d'ilot devient positivement corrélé au facteur de correction de la couverture de la strate à posteriori. Autrement dit, plus le secteur est grand, plus le sous-dénombrement du secteur détermine le facteur net de correction.

L'effet stochastique serait négligeable pour tous les secteurs, sauf les plus petits, si l'hypothèse d'indépendance autonome de Wolter (1986) était vérifiée en pratique, c'est-à-dire si chaque personne était incluse ou manquée indépendamment de toute autre personne. En fait, il est bien connu que, souvent, des familles entières sont manquées ou comptées en double. En effet, l'immeuble entier (voire même parfois l'ilot) pourrait être manqué ou compté en double au moment de l'établissement de la liste d'adresses du recensement. La non-validité de l'hypothèse d'indépendance autonome n'introduit pas de biais dans le modèle à système dual à condition que les probabilités sous-jacentes soient égales à l'intérieur des strates à posteriori. Cette non-validité peut signifier que la couverture observée pour un ilot ne concorde pas avec le rajustement estimé pour le sous-dénombrement. Cependant, à mesure qu'on se tourne vers des secteurs plus grands, l'effet stochastique diminue et est remplacé par le problème de l'hétérogénéité vraie des probabilités de saisie sous-jacentes (consulter Haines 2001, pour des précisions sur l'estimation synthétique).

personnes ayant déménagé d'après le nombre de personnes venues s'installer dans les ilots échantillonnés entre le 1^{er} avril et le moment de l'interview de l'ACB (personnes entrantes). Si l'on suppose que la population n'est pas soumise aux effets de la migration internationale, des décès, des mouvements vers les logements collectifs, etc., alors le nombre de personnes qui viennent s'établir dans les ilots doit être égal au nombre de personnes qui en sortent (personnes sortantes). Il s'agit des mêmes personnes dans la population, si ce n'est dans l'échantillon. Il est habituellement plus facile de trouver les personnes où elles sont établies, si bien que la détermination du nombre de personnes entrantes est généralement une meilleure estimation du nombre total de personnes ayant déménagé que celle du nombre de personnes sortantes.

Pour estimer la proportion de personnes ayant déménagé qui ont été recensées correctement, on applique les entrecensements pour l'ilot échantillonné et le secteur de recherche étendu, au besoin. Le nombre estimé de personnes ayant déménagé correctement recensées est alors $M'_i = (M_0/N_0) N'_i$, où M_i représente le nombre pondéré d'appartenants corrects, N'_i représente le nombre pondéré dans la population et les indices représentent le nombre total de personnes ayant déménagé (i), de personnes sortantes (o) et de personnes entrantes (i).

Si nous représentons les personnes qui n'ont pas déménagé par l'indice n , le taux global de couverture devient

$$\frac{N_{11}}{N_1} = \frac{\hat{M}_n + \hat{M}'_i}{\hat{N}_n + \hat{N}'_i}$$

L'effet de la procédure C revient à augmenter la probabilité effective de saisie dans l'enquête des personnes qui ont déménagé, donc d'augmenter l'homogénéité d'inclusion dans l'enquête par rapport à la situation de démenagement (c'est-à-dire personne ayant déménagé/n'ayant pas déménagé) (Griffin 2000).

La non-réponse et la réponse partielle peuvent survenir à diverses étapes. L'objectif du processus de dépistage des données manquantes est d'améliorer l'estimation du nombre de personnes correctement comptées (à partir de l'échantillon E) ou l'estimation du ratio de couverture (à partir de l'échantillon P). En établissant les procédures de dépistage des données manquantes, nous choisissons des méthodes qui appuient les hypothèses qui sous-tendent le DSE. À compter de l'EPC de 1990, le U.S. Census Bureau estime la probabilité qu'un enregistrement de non-réponse soit correct au lieu d'attribuer une classification « zéro/un » (Schenker 1988; Belin 1993). L'utilisation de ces méthodes pour l'ACB est décrite dans Cantwell et Ikeda (dans le présent volume).

d'interview (information de dernier ressort uniquement) au recensement.

4.5 Homogénéité à l'intérieur des strates

Le DSE exige que les probabilités de saisie soient indépendantes pour tous les individus compris dans les domaines d'estimation appelés strates à postertori. On s'approche de cette condition en rendant les strates à postertori aussi homogènes que possible en ce qui concerne les probabilités de saisie au recensement, puis en s'efforçant d'obtenir des probabilités d'inclusion aussi uniformes que possible pour l'enquête.

La répartition de la population en un grand nombre de strates à postertori relativement petites peut augmenter l'homogénéité à l'intérieur des strates. Cependant, pour les petites strates, la variance d'échantillonnage et le biais dû au ratio peuvent être importants. Le biais dû au ratio découle du fait que le DSE est intrinsèquement un estimateur par le ratio. Ce biais a tendance à diminuer à mesure que la taille de la strate à postertori augmente. En outre, la façon dont nous traitons les personnes qui ont déménagé ajoute un ratio supplémentaire (voir plus loin). Par conséquent, nous avons créé des strates à postertori ayant une taille minimale prévue d'échantillon de 100.

Pour l'ACE, nous procédons à la stratification à postertori en nous appuyant sur les variables suivantes :

1. Race / origine hispanique (7)
2. Âge / sexe (7)
3. Mode d'occupation du logement (2)
4. Taille de la région métropolitaine et type de secteur de recensement (4)
5. Taux de réponse (2)
6. Région (4)

où le chiffre entre parenthèses indique le nombre de catégories. Des précisions supplémentaires sur les strates à postertori figurent dans Haines (2001).

Les différences de couverture entre les groupes ethniques sont bien décrites. (Voir, par exemple, Robinson, Ahmed, Das Gupta et Woodrow 1993; Hogan 1993). Les différences sociales, culturelles, linguistiques et économiques peuvent faire en sorte que divers groupes ethniques réagissent différemment aux procédures de recensement.

L'analyse démographique et les enquêtes de couverture antérieures ont montré que les proportions de personnes non recensées varient selon le groupe d'âge et que le profil n'est pas le même pour les hommes que pour les femmes. Les jeunes adultes représentent l'aspect le plus important de ce profil (Robinson et coll. 1993.)

L'importance du mode d'occupation du logement a été évaluée pour la première fois après le Recensement de 1980, puis les résultats ont été appliqués lors de la stratification à postertori de 1990. Les personnes qui vivent dans un logement dont elles sont propriétaires sont moins mobiles. Elles pourraient estimer que leur enjeu dans la

collectivité est plus important, donc, être plus influencées par le programme de sensibilisation au recensement. De toute évidence, la taille de la région métropolitaine a une incidence sur les profils de logement et est corrélée à la façon dont le Censur Bureau établit ses listes d'adresses. La variable combinée « taille de la région métropolitaine et type de secteur de recensement » isole des différences dans la couverture des unités de logement. En outre, elle pourrait mesurer certains aspects de l'isolement social et économique.

Le taux de réponse au recensement mesure la collaboration du public au recensement, un prédicteur important de la couverture. Il mesure aussi directement la proportion du décomptement qui doit être faite lors du suivi pour la non-réponse au recensement. L'une des difficultés que pose cette variable est que les secteurs de recensement du pays ne font pas tous partie de l'univers du retour par la poste. Une faible proportion du recensement est réalisée par interview directe et, naturellement, n'a pas de « taux de retour du questionnaire ». Nous avons choisi de regrouper ces secteurs avec ceux pour lesquels le taux de réponse par la poste est « élevé ».

La région de recensement reflète, entre autres, d'importants différences en ce qui a trait aux profits, d'établissement et au stock de logements. Les « demeures de riches bourgeois » sont plus courantes dans le nord-est, tandis que les maisons mobiles le sont plus dans le sud. De toute évidence, la classification croisée complète peut produire des cellules très petites. L'ensemble maximal de strates à postertori permis par le plan d'échantillonnage était de 448. En fait, après regroupement des petites cellules, il restait 416 strates à postertori.

4.6 Traitement des personnes ayant déménagé

Les personnes qui déménagent entre la date du recensement et le moment de l'interview de l'enquête posent des difficultés lors de la conception d'un DSE pour l'application de recensement. Premièrement, les personnes qui déménagent sont plus susceptibles de ne pas être dénombrées le jour du recensement et au moment de l'enquête. Deuxièmement, si le « lieu de résidence habituel » d'une personne n'est pas le même au moment de l'enquête qu'au moment du recensement, il faut décider à quel endroit échantillonner cette personne.

Lors de l'EPC de 1990, les personnes ayant déménagé ont été échantillonnées à l'endroit où elles vivaient au moment de l'interview de l'enquête. Puis, on a recherché les enregistrements de recensement au lieu de résidence habituel le 1^{er} avril, et uniquement à cette date. Cette procédure a été baptisée procédure B (Marks 1979). Cette approche oblige à coder l'adresse d'après la géographie correcte le jour du recensement, puis à procéder à l'appariement. Ces activités sont complexes et demandent beaucoup de temps.

Pour l'ACE, nous avons utilisé une procédure différente baptisée procédure C. Nous avons estimé le nombre de

L'appariement. L'information au sujet des personnes qui avaient déménagé a été recueillie auprès des résidents courants. Lors des enquêtes de 1980 et de 1990, les personnes ayant déménagé ont été interviewées à leur résidence au moment de l'interview de l'EPIC. Il a ensuite été nécessaire de coder le lieu de résidence correct le jour du recensement avant de commencer l'appariement. Cette procédure s'est avérée difficile, particulièrement dans les régions rurales. L'appariement des personnes ayant déménagé n'avait jamais été automatisé auparavant. Pour l'ACE, tous les appartements, y compris ceux des personnes ayant déménagé, ont été effectués dans la grille d'îlots ou dans un îlot adjacent de l'échantillon *E*, en utilisant le même ordinateur et le même système d'appariement manuel assisté par ordinateur. La modification du traitement des personnes ayant déménagé est exposée plus loin.

4.4 Rôle de la réinterview après l'appariement

Certains cas sont de nouveau examinés sur le terrain afin de recueillir des renseignements supplémentaires après qu'on ait achevé l'appariement initial. Cette réinterview après l'appariement est souvent appelée « interview de suivi ».

À l'instar de toutes les activités de l'EPIC, le processus d'interview de suivi doit s'inscrire dans le cadre général du DSE. Plus précisément, il doit rendre compte des éléments suivants :

1. réponse applicable, unique et correcte;
2. indépendance entre les probabilités d'inclusion dans le recensement et dans l'enquête;
3. équilibre entre les concepts des échantillons *P* et *E*;
4. secteur de recherche et règles d'appariement d'endroit unique;
5. traitement des données manquantes.

Le suivi n'est utile que s'il fournit des réponses plus exactes et plus cohérentes. Obtenir simplement une réponse différente n'est pas une justification. Puisque le suivi a lieu à une date plus éloignée de la date du recensement que l'interview initiale, il est plus difficile d'obtenir des réponses exactes. Il en est également ainsi pour le suivi de l'échantillon *E* et celui de l'échantillon *P*. Pour obtenir de meilleures réponses, le suivi doit être réalisé avec de meilleurs ressources, par exemple 1) de meilleurs répondants (ménage *c*, répondant par procuration), 2) un intervieweur mieux formé, mieux supervisé ou dont la qualité du travail est mieux contrôlée ou 3) de meilleures questions ou procédures d'interview.

La période de collecte des données du recensement s'étend de la mi-mars au milieu de l'été. Étant donné la très grande échelle de l'opération, peu d'attention est accordée à la vérification du fait que les personnes recensées

résidaient dans le logement le 1^{er} avril. La réinterview d'assurance de la qualité pour éviter la fausseté est minimale. Parce que les intervieweurs sont mieux formés et mieux supervisés, et que le questionnaire est plus complet, l'interview de suivi de l'ACE peut, en général, produire des renseignements plus exacts sur le lieu de résidence que ceux recueillis durant le recensement proprement dit. Donc, tous les cas non appariés de l'échantillon *E* ont fait l'objet d'un suivi.

Le suivi peut, toutefois, compromettre la condition d'indépendance. Si tous les cas faisaient l'objet d'un suivi, l'indépendance ne serait pas nécessairement compromise. Cependant, les cas appariés durant le premier exercice d'appariement font rarement l'objet d'un suivi, car l'épuisement des ressources disponibles serait trop important. Au contraire, seuls les cas de non-appariement ou « d'appariement possible » sont habituellement sélectionnés pour le suivi, ce qui pourrait donner lieu à une dépendance opérationnelle.

Le biais dû au suivi peut se produire même si l'interview est réalisée correctement, car le suivi peut modifier sélectivement l'« endroit correct » défini pour les non-appariements, mais non pour les appariements. Si les opérations de suivi aboutissent à une non-interview, un biais supplémentaire peut être introduit selon les modèles de données manquantes appliqués à ces cas.

Pour choisir les cas qui feront l'objet d'un suivi, il faut trouver un juste équilibre entre l'obtention d'information exacte et cohérente, d'une part, et la nécessité de satisfaire à la condition d'indépendance, d'autre part. Dans le cas de l'échantillon *P*, seuls ont été suivis les cas où il était probable d'obtenir de meilleurs renseignements. Les cas désignés pour le suivi incluaient :

1. les appartements possibles, puisque, grâce à l'information disponible, les intervieweurs permettent de résoudre la situation;
2. les cas d'interviews initiales réalisées auprès d'un répondant par procuration ne faisant pas partie du ménage donnant lieu à un non-appariement, puisque, comme on n'a pas parlé à un membre du ménage, on a des raisons de douter de l'exactitude des renseignements;
3. les cas non appariés où, pour la même unité de logement, le recensement fait état d'une famille et l'ACE, d'une autre; pour assurer la concordance de la déclaration de l'adresse le jour du recensement entre l'échantillon *P* et l'échantillon *E*, ces cas sont suivis ensemble;
4. les non-appariements de ménages partiels.

Les cas appariés et certains autres cas non appariés n'ont généralement pas fait l'objet d'un suivi. Par exemple, dans le cas de l'ACE, on n'a pas fait le suivi des cas de ménage entier non appariés pour lesquels l'unité avait été manquante, déclarée vacante ou pour laquelle on n'avait pu obtenir

d'abord si un cas était acceptable pour l'appariement, puis seulement après de faire une tentative d'appariement, la logique étant de ne pas essayer de découvrir un appariement à moins qu'on ne soit certain que, si aucun appariement n'a lieu, la personne n'a pas été recensée.

Avant de procéder à l'appariement, on a examiné les enregistrements de l'échantillon *P* pour déterminer s'ils satisfaisaient aux critères :

- 1) d'applicabilité;
- 2) d'unicité;
- 3) de complétude;
- 4) d'exacritude géographique

L'échantillon de l'ACE ne contenait aucun enregistreur manifestement fantaisiste. À cet égard, le recours à l'interview sur place assistée par ordinateur (IPAO) offre une protection importante. L'instrument d'IPAO rend la falsification difficile grâce au « marquage de la date et de l'heure » de l'interview et à l'enregistrement de chaque trappe. Nous avons mis en place un processus d'assurance de la qualité visant à réduire au minimum d'autres pratiques d'interview négligentes ou malhonnêtes durant l'ACE. En outre, une exception importante à la règle du « pas de suivi » a été faite pour les cas où la falsification était possible, c'est-à-dire ceux où il n'y a d'appariement pour aucun membre du ménage, ce qui implique une falsification éventuelle.

Les enregistrements hors champ, comme les logements collectifs, ont été délistés et éliminés. Les éventuels enregistrements d'enquête en double ont également été délistés et éliminés (unicité). Enfin, si l'interview ne satisfaisait pas aux normes minimales, le cas a été converti en un cas de non-réponse et soumis plus tard à l'imputation.

4.2 Déclaration uniforme du lieu de résidence

Pour établir correctement le nombre de personnes dans les deux systèmes, nous devons déterminer si les personnes comprises dans l'échantillon *P* ont été ou non recensées correctement. Il faut pour cela rechercher les enregistrements de recensement corrects dans le secteur où la personne aurait dû être recensée.

Une même définition de l'exacritude géographique doit être appliquée pour déterminer si un enregistrement (dans l'échantillon *E*) est correct ou si une personne (dans l'échantillon *P*) a été recensée correctement. En cas de non-concordance entre ces concepts, on parle d'« erreur d'équilibrage ».

Plus précisément, nous devons utiliser la même définition de l'« endroit correct » et le même secteur de recherche autour de l'endroit correct. Les erreurs peuvent donner lieu aussi bien à des non-appariements qu'à des appariements erronés. Les difficultés proviennent principalement de deux sources. Premièrement, des réponses par procuration sont acceptées dans le cas de l'échantillon *P*

L'objet de l'appariement est de déterminer si une personne interviewée en tant que membre de l'échantillon *P* a aussi été dénombrée au recensement dans le secteur de recherche défini. Aujourd'hui, la plupart des appariements sont informatisés. Le système produit des appariements, des appariements possibles et des cas non appariés. Des essais répétés ont montré que les enregistrements appariés par l'ordinateur sont presque certainement correctement couplés (Belin 1993). Presque tous les appariements manuels sont maintenant assistés par ordinateur et, dans la plupart des cas, n'incluent aucune paperasserie. Ce nouveau système facilite la recherche, y compris celle des enregistrements en double. Il limite le nombre de codes que les commis peuvent appliquer à ceux qui sont appropriés pour la situation. L'absence presque totale de paperasserie a éliminé les cas de perte ou de classement erroné des questionnaires de l'ACE.

4.3 Appariement exact

Les commis de premier niveau ont été appuyés par une équipe de 46 techniciens. La formation de ces techniciens a débuté en septembre 1999. Ils ont eux-mêmes été appuyés par une équipe de sept analystes permanents, dont la plupart avaient effectué des appariements pendant de nombreuses années. Chaque niveau d'appariement joue le rôle d'assurance de la qualité pour le niveau précédent. En outre, chaque niveau peut adresser les cas problématiques au niveau directement supérieur. Tous les appariements ont été effectués au même endroit par un seul groupe d'employés. Lors des enquêtes de 1980 et de 1990, les opérations d'appariement ont eu lieu à trois et à sept endroits, respectivement.

L'application des procédures de l'ACE pour les personnes ayant démenagé a aussi simplifié beaucoup

peuvent être traité de façon analogue aux cas d'imputations de personne entière au recensement, c'est-à-dire en remplaçant $(C - II) - ETR$ dans l'équation 6. L'exclusion des ETR n'a pas d'effet sur le DSE de la population réelle si le nombre d'appartements est réduit proportionnellement au nombre d'enregistrements de recensement corrects. Autrement dit, on suppose que la probabilité d'inclusion dans l'ACE, il s'agit, naturellement, de l'hypothèse d'indépendance habituelle du système dual. (Voir Hogan 2001, pour la théorie sous-jacente.) Bien qu'on ait dénombré 2,3 millions d'ETR au Recensement de 2000, l'analyse des résultats de l'ACE faite par Raglin (2002) montre que l'effet sur les résultats finaux du DSE est banal.

Dans les situations où le nombre d'imputations de personne entière (III) est faible, $(CE / N_p - I)$ représente une mesure du surdénombrement brut au recensement. Cependant, cette mesure est une fonction des définitions opérationnelles de « correctement recensé » adoptées pour établir le plan de l'enquête d'évaluation de la couverture. Les définitions adoptées pour produire une bonne mesure de la couverture nette, particulièrement en ce qui concerne la complétude et l'exactitude géographique, peuvent différer de celles qui sont les plus appropriées pour étudier la qualité des opérations de recensement sur le terrain. Quoi qu'il en soit, le Recensement de 2000 comptait 5,8 millions d'imputations de personne entière, dont 1,2 million ont eu lieu pour des unités de logement où l'intervieweur n'avait même pas pu obtenir le nombre de résidents (voir tableau 1 dans Nash 2001 et la page ii de Wetogran et Cresce 2001).

4. DÉTERMINATION DE LA PROPORTION DE PERSONNES CORRECTEMENT RECENSÉES

Après avoir défini l'ensemble de personnes correctement recensées, l'étape suivante du DSE consiste à estimer le taux de couverture du recensement, N_1 / N_p . Conceptuellement, pour estimer le taux, il faut (1) tirer un échantillon de personnes, (2) déterminer si ces personnes auraient dû être dénombrées lors du recensement et (3) déterminer si elles ont, effectivement, été recensées correctement, en se servant des mêmes définitions que celles utilisées pour mesurer N_1 . Si nous pouvons sélectionner un échantillon sans biais de personnes qui auraient dû être recensées et que nous pouvons déterminer si elles ont effectivement été recensées correctement (inclues dans le recensement), alors le DSE produira des estimations asymptotiquement correctes, si chaque étape peut être approximativement correcte, les résultats s'approcheront d'une estimation sans biais. La première étape du processus consiste, normalement, à tirer un échantillon aléatoire. On utilise pour cela dans le cas de l'ACE le même ensemble de grappes d'îlots que celui utilisé pour définir l'échantillon E.

1. l'indépendance des opérations;
2. la déclaration cohérente du lieu de résidence;
3. l'appartenance exact;
4. l'homogénéité à l'intérieur des strates a posteriori.

4.1 Indépendance opérationnelle

Bien qu'il s'agisse de l'hypothèse dont l'approximation est la plus facile, l'indépendance opérationnelle demande néanmoins de la vigilance. Lors du Recensement de 2000, le tirage de l'échantillon de l'ACE et l'établissement de la liste d'unités de logement ont eu lieu avant la livraison des questionnaires de recensement. Même si le contact personnel est minime, chez certaines personnes, le fait de figurer sur une liste établie pour une enquête peut modifier l'attitude à l'égard du recensement. Pour les unités de logement énumérées indépendamment et couplées à une adresse de recensement pour laquelle un questionnaire de recensement avait été rempli, on a pu procéder tôt aux interviews téléphoniques. Celles-ci ont eu lieu alors que le suivi pour la non-réponse au recensement se poursuivait encore dans le secteur de recensement. Dans certains cas, l'interview sur place a eu lieu en même temps que l'interview visant à améliorer la couverture du recensement. Par conséquent, une certaine contamination pourrait avoir eu lieu. On s'est efforcé par tous les moyens d'éviter qu'une même équipe d'intervieweurs travaille dans le même secteur durant le recensement et durant l'ACE et d'empêcher le partage d'information. Néanmoins, certaines personnes pourraient réagir différemment à l'enquête selon qu'elles ont été recensées, par exemple, par un recenseur très aimable ou très révéche. D'autres pourraient penser qu'elles sont tenues de fournir l'information une fois, mais non deux.

Il faut aussi assurer l'indépendance des procédures administratives. Il arrive que les définitions de la « non-réponse » ou de l'« information suffisante » soient appliquées différemment aux enregistrements apparus et non apparus de l'échantillon P. Dans l'ACE, pour éviter d'introduire inutilement une dépendance opérationnelle, on a forcé le système de traitement des données à décider

un grand flot. Pour l'ACE, on a exigé que les intervieweurs trouvent au moins trois répondants bien informés avant de coder un enregistré comme étant celui d'une personne fictive. Cependant, puisque la personne peut avoir vécu ailleurs dans l'ilot, il est parfois difficile de procéder à ce codage.

L'obligation d'accepter des réponses par procuration pour vérifier nombre d'enregistrements de recensement est une source importante d'erreur. Si la personne qui répond par procuration déclare un lieu de résidence « correct » différent de celui que la personne proprement dite déclarerait, l'enregistrement en question pourrait être mal codé, puisque l'exigence d'un lieu de résidence « correct » unique ne serait pas respectée. Pour l'ACE, l'interview par ménage a été utilisée pour les ménages qui avaient déménagé entre le moment du recensement et le moment de l'interview de l'enquête. Même au sein d'un ménage, l'opinion quant au lieu de résidence « correct » d'une personne le jour du recensement peut varier d'un membre à l'autre. Les répondants par procuration, aussi bien faisant partie du ménage que n'en faisant pas partie, ont été la cause d'un grand nombre d'erreurs de déclaration du lieu de résidence lors de l'ACE et, donc, de la sous-estimation de l'erreur au recensement.

Après l'estimation des cas de données manquantes et la pondération de l'échantillon, nous pouvons estimer le nombre de personnes dénombrées correctement au recensement comme étant

$$N_{+1} = (C - II) \frac{CE}{N_e} \quad (6)$$

où
 C = nombre total d'enregistrements au recensement, y compris les enregistrements imputés, les enregistrements en double, les enregistrements fictifs, etc. (c'est-à-dire le chiffre du recensement);
 II = nombre d'imputations de personne entière au recensement;
 CE = estimation pondérée du nombre d'enregistrements de recensement applicables, uniques, complets et corrects;
 N_e = estimation pondérée du total d'après l'échantillon E , y compris les enregistrements en double, les enregistrements fictifs, etc.

Parfois, à cause d'erreurs de traitement des données ou de contraintes de temps, il arrive qu'un groupe d'enregistrements du recensement soit exclu à la fois du traitement de l'échantillon E et du processus de recherche et d'appariement. Par conséquent, alors que ces enregistrements peuvent être traités dans les délais prévus pour être inclus dans les résultats officiels du recensement, ils arrivent trop tard pour être inclus dans le traitement des données d'évaluation de la couverture. Ces cas, parfois appelés « enregistrements tardifs de recensement » (ETR),

personne était réellement âgée de 19 ans, mais a été comprise au recensement comme ayant 17 ans, elle est malgré tout considérée comme correctement incluse. Ce point est examiné à la section 5.2.

Pour estimer le nombre de personnes correctement incluses dans le recensement, on doit tirer un échantillon de l'ensemble des enregistrements de recensement définis par des données. Cet échantillon est appelé échantillon d'enregistrement du recensement (ou E). Les imputations de la base de sondage dont est tiré l'échantillon E . Pour maximiser la corrélation avec l'échantillon de population (voir plus loin), pour l'ACE, on commence par définir un ensemble de zones d'échantillonnage qui correspondent à un flot unique ou à un groupe d'îlots contigus, ou grappe d'îlots. Si un îlot est échantillonné, tous les enregistrements de recensement codés comme correspondant à cet îlot, même incorrectement, font partie de l'échantillon. Les îlots qui contiennent un grand nombre d'enregistrements d'unité de logement peuvent faire l'objet d'un sous-échantillonnage.

Les enregistrements de l'échantillon E sont vérifiés pour confirmer leur complétude. Seuls les enregistrements qui satisfont aux exigences minimales de complétude peuvent être considérés comme des enregistrements de recensement corrects. Ensuite, on procède à une recherche des enregistrements dans le secteur de recherche pour voir si la personne a été comptée plus d'une fois dans l'îlot échantillonné (unicité). La recherche des enregistrements en double est réalisée par appariement manuel assisté par ordinateur. Lorsque plus d'un enregistrement est repéré pour une même personne, les enregistrements supplémentaires reçoivent le code d'enregistrement en double. L'applicabilité et l'exacitude géographique ne peuvent être vérifiées d'après les enregistrements du recensement uniquement et nécessitent une interview supplémentaire. Si l'intervieweur dépiste un membre du ménage ou un répondant acceptable qui peut confirmer l'existence de la personne et que l'endroit en question était le lieu de résidence habituel de cette personne le 1^{er} avril, l'enregistrement de recensement est considéré comme correct. Si le répondant déclare que la personne ne vivait pas dans l'îlot ni dans le secteur de recherche le 1^{er} avril, l'enregistrement est considéré comme étant incorrect. Cette situation peut survenir si une personne a répondu au recensement, mais a déménagé avant le 1^{er} avril, si une personne a emménagé après le 1^{er} avril, mais a été dénombrée au moment de l'opération de suivi pour la non-réponse au recensement ou qu'un parent déclare incorrectement qu'un étudiant collégial ou universitaire vivait au domicile parental.

Les intervieweurs peuvent déterminer que la personne n'a jamais existé ou qu'elle n'a jamais été associée à l'îlot en question. Ces enregistrements sont considérés comme incorrects. Il est difficile, dans certains cas, de prouver qu'une « personne » n'était pas réelle, particulièrement dans

impose des conditions de « définition d'un enrégistrement-personne par des données ». Pour le Recensement de 2000, deux caractéristiques étaient nécessaires pour définir un enrégistrement, le nom comptant comme une caractéristique. La zone du nom doit comporter au moins trois caractères pour les zones combinées du prénom et du nom de famille. Les caractéristiques incluses dans le dénombrement sont le lien avec le chef du ménage, le sexe, la race, l'origine hispanique et l'âge ou l'année de naissance (Childers 2001).

Lorsqu'un enrégistrement ne satisfait pas à ces exigences, à l'étape du traitement des données du recensement, on lui substitue (impute) un enrégistrement correctement défini. Comme le module de traitement reconnaît toutes ces imputations de personne entière, leur nombre est connu et ne doit pas être estimé. Habituellement, le nombre d'imputations de personne entière est dénoté II, pour « information insuffisante ».

En outre, certains enrégistrement-personne sont acceptables pour l'étape du traitement des données, mais insuffisants l'application du DSE. Ce groupe inclut les enrégistrement contenant des données raisonnablement complètes, mais où ne figure pas le nom de la personne. Un appareil exact ou une interview supplémentaire est impossible pour ces cas. Pour l'ACE de 2000, l'« information suffisante pour l'appareil » s'entend du nom complet et de deux caractéristiques (Childers 2001).

« Exactitude géographique » signifie que les personnes ont été recensées à l'endroit où elles devaient l'être. Les personnes dénombrées à l'extérieur de ce ou ces secteurs de recherche définis sont comptées dans le recensement, mais ne sont pas incluses correctement dans celui-ci. Le secteur en question doit faire l'objet d'une recherche durant le processus d'appareil, ainsi que d'une recherche des adresses dans le secteur de recherche enregistré de recensement en double. À mesure que le nombre d'adresses dans le secteur de recherche augmente, la complexité de l'appareil et la probabilité d'une erreur d'appareil augmentent aussi. Cet accroissement de la complexité et du niveau possible d'erreur aura une incidence sur l'appareil entre les enrégistrement de l'enquête et ceux du recensement, ainsi que sur la recherche d'enrégistrement de recensement en double. La probabilité qu'on manque un appareil réel est d'autant plus forte que le nombre d'adresses visées par la recherche est élevé. Et, fait tout aussi important, la probabilité d'un appareil incorrect augmente. Par exemple, la probabilité de trouver deux personnes ayant le même nom et le même âge vivant dans le même îlot est faible. Par contre la probabilité de trouver deux personnes ayant ces caractéristiques dans une grande ville est considérable.

Deux dimensions doivent être définies pour rendre un secteur de recherche opérationnel, à savoir 1) l'endroit correct et 2) le secteur de recherche autour de l'endroit correct. L'« endroit correct » précise où, en vertu des règles de résidence du DSE, la personne doit être incluse dans le correct.

La personne est incluse correctement dans le recensement si elle l'est à l'endroit qu'elle considérait, au moment de l'interview de l'enquête, comme son lieu de résidence habituel au 1^{er} avril.

Dans l'ensemble, cette définition est conforme aux règles du recensement. Cependant, elle tient compte explicitement du fait que le concept de « lieu de résidence habituel » est, dans une certaine mesure, subjectif. À cause de cette subjectivité, le lieu que la personne considère comme ayant été son lieu de résidence habituel (au 1^{er} avril) peut avoir changé au moment de l'interview de l'enquête. Cette situation, en soi, ne biaise pas le DSE. Cependant, elle nécessite que la déclaration de l'« endroit correct » soit cohérente.

La deuxième dimension de l'exactitude géographique est le secteur de recherche autour de l'endroit correct. Le concept de secteur de recherche a pour but de tenir compte de l'affectation incorrecte de résidents à une zone géographique particulière lors du recensement ou de l'enquête. Son application réduit la variance et peut, dans certaines circonstances, réduire aussi le biais.

La définition qui suit a été utilisée pour l'ACE :

Une personne a été recensée correctement si elle a été comptée dans la grappe d'îlots contenant son lieu de résidence habituel; ou si elle a été incluse lors du recensement dans l'unité de logement où elle réside habituellement et que l'unité de logement a été incluse dans un îlot adjacent à la grappe d'îlots correcte.

Un aspect important de ce plan d'enquête est que les enrégistrement correspondant à des personnes recensées au « mauvais » endroit doivent être considérés comme erreurs. Donc, une personne comptée une seule fois, mais au mauvais endroit devrait être catégorisée, en moyenne, comme contribuant à l'endroit correct. Cette méthode évite de (une omission) à l'endroit correct. Cette méthode évite de devoir procéder à une recherche étendue d'enrégistrement en double éventuels, mais exige qu'on détermine lors des interviews sur le terrain un endroit correct unique pour chaque personne.

La définition de « correctement recensé » ne dépend pas de l'exactitude de la classification. Par exemple, si une

Par exemple, f peut définir tous les Asiatiques de 0 à 17 ans vivant dans une unité de logement occupée par son propriétaire, tandis que k peut définir Orange County, Californie, et h peut définir les filles de 11 ans. Ce calcul produit une estimation pour une petite région et pour un petit groupe, mais il peut donner lieu à des fractions. Habituellement, l'utilisateur des données de recensement préfère des enregistrements portant sur des personnes entières. On recourt, pour le recensement des Etats-Unis, à un arrondissement contrôlé et à l'imputation d'enregistrements-personne pour créer un nombre entier rend les données plus acceptables.

3. DÉTERMINATION DU NOMBRE D'ENREGISTREMENTS DE RECENSEMENT CORRECTS

3.1 Définition et dénombrement des enregistrements corrects et erronés

La première étape de l'application de l'équation 1 consiste à définir et à estimer l'ensemble d'individus comptés « correctement » lors du recensement. Dans ce contexte, le terme « correctement » a les quatre dimensions suivantes :

- 1. applicabilité;
- 2. unicité;
- 3. complétude;
- 4. exactitude géographique.

« Applicabilité » signifie que la personne devrait être incluse dans le recensement. Les personnes décédées avant le jour du recensement ou nées après celui-ci (1^{er} avril aux Etats-Unis) ne font pas partie de la population (univers) qui doit être mesurée. Par conséquent, les enregistrements ayant trait à des « personnes » fictives, des touristes ou des animaux sont hors du champ.

« Unicité » a trait au fait que nous voulons déterminer le nombre de personnes incluses dans le recensement, et non le nombre d'enregistrements de recensement. Si plus d'un enregistrement se rapporte à une même personne, il faut réduire le nombre d'enregistrements afin d'utiliser le DSE. « Complétude » signifie que l'enregistrement de recensement doit être suffisant pour identifier une personne unique. Si les renseignements d'identification sont insuffisants, nous ne pouvons déterminer si la personne a été incluse de façon applicable et unique dans le recensement, ni si elle a aussi été incluse dans l'enquête.

Bien que la complétude soit nécessaire pour utiliser le DES, le chiffre du recensement de recensement incomplets, d'autres enregistrements de recensement, pour les opérations de recensement, on Normalement,

recensement par le rapport du nombre compté durant l'enquête au nombre compté dans les deux systèmes (c'est-à-dire l'inverse du taux de couverture du recensement, tel que mesuré d'après l'enquête).

Le DSE donnera une estimation directe de la population de la classe f , ainsi que toute somme des classes. La classe f pourrait être la population de ménages d'un état, d'un district, d'un groupe ethnique ou, peut-être, d'un groupe ethnique à l'intérieur d'un Etat.

Souvent, les conditions à remplir pour estimer de petites populations ou des populations locales, par exemple, par âge selon le sexe, la race, ou la ville, excèdent même la capacité d'un très grand échantillon. Pour produire ce genre d'estimation, on combine le DSE à une hypothèse synthétique pour produire des estimations pour des régions géographiques plus petites que celles définies par le domaine f . L'estimateur synthétique suppose qu'une portion, ou ratio, mesurée à un niveau agréé s'applique uniformément à tous les sous-groupes (Gonzalez 1973; Gonzalez et Hoza 1978). En utilisant une hypothèse synthétique, nous écrivons

$$N_{js}^{fkh} = CCF_f C_{jkh} \tag{2}$$

$$CCF_f = \frac{N_f}{C_f} \tag{3}$$

N_{js}^{fkh} est la population estimée du domaine f , disponible au niveau de géographie k et de la sous-classe h est la mesure (habituellement le chiffre du recensement) de la population du domaine f disponible au niveau de géographie k et de la sous-classe démographique h ;

CCF_f est le facteur net de correction de la couverture;

N_f est la mesure (habituellement le chiffre du recensement) de la population du domaine f ;

C_{jkh} est la mesure (habituellement le chiffre du recensement) de la population du domaine f disponible au niveau de géographie k et de la sous-classe démographique h .

$$C_f = \sum_k \sum_h C_{jkh} \tag{4}$$

$$N_s^k = \sum_h \sum_f CCF_f C_{jkh} \tag{5}$$

C_f ne doit pas nécessairement être égal au nombre de personnes correctement incluses dans le recensement (N_{+1}). N_{+1} est estimé d'après des données d'échantillon et n'est pas disponible pour toutes les petites régions. C est normalement le chiffre de recensement, y compris les imputations et les inclusions erronées (enregistrements en double, etc.).

La sommation sur le groupe f et la sous-classe h donne une mesure de la population de la région géographique donnée k :

L'évaluation de l'exacritude et de la couverture : Théorie et conception

HOWARD HOGAN¹

RÉSUMÉ

Le présent article porte à la fois sur la question générale de la conception d'une enquête postcensitaire et sur la façon dont cette question générale a été traitée par le U.S. Census Bureau lors de la mesure de la couverture planifiée dans le cadre du Recensement de 2000. Il vise à établir le lien entre les concepts fondamentaux de l'estimateur à système dual et les questions de la définition et du dénombrement des enrégistrement de recensement corrects, du dénombrement des omissions au recensement, de l'indépendance opérationnelle, de la déclaration du lieu de résidence, ainsi que du rôle de la réinterview après apparement. Y sont discutés des problèmes d'estimation, comme le traitement des personnes ayant des défauts de données manquantes et l'estimation synthétique du chiffre de population local corrigé. Sont aussi exposés les défauts de conception de l'évaluation de la couverture du Recensement de 2000.

MOTS CLÉS : Estimation selon le système dual; correction des données du recensement; sous-dénombrement.

1. INTRODUCTION

Le U.S. Census Bureau s'est efforcé de corriger les

chiffres de population initiaux du Recensement de 2000 pour tenir compte du sous-dénombrement net mesuré (U.S. Census Bureau 2000). Cette correction devait se fonder sur

l'évaluation de l'exacritude et de la couverture (ACE pour *Accuracy and Coverage Evaluation*). Cette dernière est une

enquête postcensitaire fondée sur l'estimateur à système

dual (DSE pour *Dual System Estimator*). Quoi qu'en

apparece bien conçue et bien exécutée, l'ACE a produit

des estimations initiales très imparfaites. Elle a donné une

estimation du sous-dénombrement net de 3,3 millions (é.-l. :

378 000), résultat qui diffère fortement de l'estimation de

340 000 seulement fondée sur l'analyse démographique

(Robinson 2001), ainsi que d'une estimation d'enquête

révisée ultérieurement donnant un surdénombrement de

1,3 million (é.-l. : 542 000) (U.S. Census Bureau 2003).

Le présent article traite à la fois de la question générale

de la conception d'une enquête postcensitaire (EPC) et de

la façon dont cette question a été abordée par le

U.S. Census Bureau lors de la planification de l'ACE. Le

cas échéant, il décrit les situations où les hypothèses

sous-tendant la conception de l'enquête de 2000 n'ont pas

été vérifiées. Dans tout l'exposé, les expressions estimateur

à système dual (DSE) et enquête postcensitaire (EPC) sont

utilisées lors de la discussion d'une question générale et

l'expression évaluation de l'exacritude et de la couverture

(ACE) est employée lors de la discussion de détails parti-

culiers au plan de l'enquête amériaine de 2000. La section

suivante donne la définition du modèle à système dual

appliquée à la mesure de la couverture du recensement. La

section 3 porte sur la définition et le dénombrement des

enrégistrement de recensement corrects et erronés. La

section 4 décrit les problèmes que posent la définition et le

dénombrement des omissions. La section 5 traite de l'esti-

mation de petites régions. L'article se termine par la

discussion de certains problèmes qu'il a fallu surmonter lors

de la mise en œuvre de l'ACE et l'énoncé de certaines

conclusions.

2. LE MODÈLE D'ESTIMATION À SYSTÈME DUAL

L'utilisation du modèle à système dual est répandue pour

évaluer la complétude de l'enregistrement des événements

démographiques (Sekar et Deming 1949; Marks, Selizer et

Wolter 1986; U.S. Bureau of the Census 1985). L'application du

modèle à système dual dans le contexte du Recensement de

1990, y compris la question de la correction des données de

recensement, est décrite dans Hogan (1992, 1993).

L'estimateur type de Petersen (1896), de Sekar-Deming

ou à système dual (DSE) peut s'exprimer sous la forme :

$$\hat{N}_{++} = N_{+1} (N_{1+} / N_{11}) \quad (1)$$

où

N_{11}

est le nombre de personnes comptées à la fois

durant le recensement et durant l'enquête;

N_{+1}

est le nombre de personnes comptées durant

l'enquête;

N_{+}

est le nombre total de personnes.

Autrement dit, on obtient l'estimation de la population

totale en multipliant le nombre enregistré lors du

Nous remercions le rédacteur adjoint, Fritz Scheuren, de ses observations intéressantes au sujet de notre réplique.

RÉFÉRENCES ADDITIONNELLES

- COX, L.H. (1994). Méthodes de masquage de matrice pour la protection du caractère confidentiel de microdonnées. *Techniques d'enquête*, 20, 173-177.
- RAGHUNATHAN, T.E., REITER, J.P. et RUBIN, D.B. (2002). Multiple imputation for statistical disclosure limitation. Rapport technique, Department of Biostatistics, University of Michigan, Ann Arbor.

proportionnelle à une mesure de la taille de la grappe M_i , qui n'est pas nécessairement égale à la taille réelle de la grappe M_i . Dans ce cas, nous ne pouvons appliquer l'algorithme du cas 3 de la section 2 pour obtenir un sous-échantillon aléatoire simple.

Des travaux supplémentaires sont manifestement nécessaires en vue de développer des algorithmes permettant d'obtenir un échantillon exact, ou au moins un échantillon aléatoire stratifié. Comme l'a mentionné Fritz Scheuren lors d'une communication personnelle, ce problème est amusant, mais il y a encore beaucoup à faire.

(vi) Représentation graphique et modélisation

L'application directe des méthodes standard de représentation graphique et de modélisation de statistiques aux données d'enquête complexes peut produire des graphiques et des modèles incorrects, comme le souligne Eltinge, à cause des effets de la mise en grappe, des poids inégaux, de la stratification et d'autres caractéristiques des données d'enquête. Par ailleurs, l'application des méthodes standard aux données provenant d'un échantillon inverse simple (ou sous échantillon) est apparemment une condition que le sous échantillon soit inconditionnellement un échantillon aléatoire simple. Cependant, la taille du sous échantillon, m , est habituellement petite et l'ensemble de données sur le sous échantillon n'est donc pas informatif pour la représentation graphique ou la modélisation. On peut faire passer la taille de l'ensemble de données à gm en combinant les g sous échantillons, mais l'application des méthodes standard (par exemple les nuages de points) à l'ensemble de données combinées pourrait produire des graphiques trompeurs et des inférences incorrectes parce que les sous échantillons sont corrélés inconditionnellement. Eltinge a proposé certains moyens utiles de tenir compte de la corrélation non conditionnelle dans le contexte de l'estimation bivariate de la densité, mais beaucoup de travaux restent à faire dans le domaine de la représentation graphique et de la modélisation statistiques au moyen d'échantillons inverses multiples.

(vii) Confidentialité des microdonnées

Comme le souligne Eltinge, un des grands attraits éventuels de l'échantillonnage inverse est qu'il permet le calcul d'estimateurs ponctuels, d'erreurs types, etc. à partir du fichier de microdonnées, correspondant à de multiples sous échantillons, sans devoir connaître les poids de sondage, les identificateurs de grappe ou les identificateurs de strate. Cette caractéristique permet de réduire le risque d'identification que pose la connaissance des identificateurs de grappe, etc. Déterminer la mesure dans laquelle le fichier de données fondé sur des sous échantillons multiples permet aux utilisateurs des données de reconstruire les poids ou les identificateurs de grappe pourrait être une tâche difficile. Notons que les valeurs caractéristiques figurant dans le fichier de données provenant d'échantillons inverses sont réelles au sens de leur correspondance aux valeurs de l'échantillon complet.

Si l'échantillon complet est un échantillon en grappes PPT et que les sous échantillons sont obtenus par sélection d'un élément à partir de chaque grappe, alors on peut éviter l'identification des grappes en permettant d'abord aléatoirement les vecteurs de données à l'intérieur de chaque sous échantillon, puis en publiant les données pour les sous échantillons perméables. La méthode des EBC n'est pas influencée par la permutation des vecteurs de données à l'intérieur de chaque sous échantillon.

Il convient de souligner que la protection des renseignements confidentiels qu'offre l'ensemble de données obtenus par l'échantillonnage inverse est une étape avec probabilité n'est pas nécessairement une opération viable, à moins qu'on anticipe son utilisation à l'étape de l'élaboration du plan d'échantillonnage complet afin de permettre l'utilisation de plans de sondage inversibles. À l'heure actuelle, nous ne disposons d'aucune procédure d'échantillonnage inverse pour plusieurs plans de sondage à échantillon complet utilisés couramment. Par exemple, considérons l'échantillonnage en grappes à une étape avec probabilité

(viii) Conclusion

Comme le mentionne Hinkins, l'échantillonnage inverse n'est pas nécessairement une opération viable, à moins qu'on anticipe son utilisation à l'étape de l'élaboration du plan d'échantillonnage complet afin de permettre l'utilisation de plans de sondage inversibles. À l'heure actuelle, nous ne disposons d'aucune procédure d'échantillonnage inverse pour plusieurs plans de sondage à échantillon complet utilisés couramment. Par exemple, considérons l'échantillonnage en grappes à une étape avec probabilité n'est pas nécessairement une opération viable, à moins qu'on anticipe son utilisation à l'étape de l'élaboration du plan d'échantillonnage complet afin de permettre l'utilisation de plans de sondage inversibles. À l'heure actuelle, nous ne disposons d'aucune procédure d'échantillonnage inverse pour plusieurs plans de sondage à échantillon complet utilisés couramment. Par exemple, considérons l'échantillonnage en grappes à une étape avec probabilité

être choisi en tenant compte des paramètres importants. Mais, naturellement, pas tous.

v) Analyse de données d'enquête

Le calcul d'erreurs types valides des estimateurs des paramètres pour un ensemble de microdonnées en échantillon complet n'est pas toujours faisable dans le contexte de l'échantillonnage stratifié à plusieurs degrés sans l'identification des grappes et des strates dans le fichier des erreurs types existe dans l'ensemble de données, l'analyste pourrait ne pas disposer du logiciel pour enquête complète approprié pour bon nombre des analyses prévues, voire toutes, comme le fait remarquer Eltinge. Par ailleurs, on pourrait obtenir des erreurs types valides par la méthode des EBC en utilisant des fichiers de microdonnées contenant plusieurs sous échantillons aléatoires simples sans qu'il soit nécessaire de connaître les poids de sondage, les identificateurs de grappes, etc. En outre, comme le souligne Eltinge, les étapes de calcul supplémentaires que demande l'application de la méthode des EBC n'imposent qu'un léger fardeau de plus à l'analyste ou peuvent être absorbées dans le logiciel analytique de façon transparente pour l'analyste. Cependant, nous devons poursuivre les travaux en vue d'améliorer les logiciels standard afin de pouvoir appliquer la méthode des EBC en pratique.

Hinkins, Oh et Scheuren (1995) ont combiné les sous échantillons pour tester l'indépendance dans un tableau de contingence 2 x 2. Leur variable chi carré de Pearson est de la forme

$$X^2 = (gm) \sum_{j=1}^{l-1} \sum_{i=2}^{l-1} (p_{ijg}^{1g} - p_{ijg}^{1g} p_{+jg}^{1g} / (p_{i+g}^{1g} p_{+jg}^{1g}))^2$$

où p_{ijg}^{1g} est l'estimateur combiné en échantillonnage inverse de la proportion dans la (i, j) ième cellule P_{ij}^{1g} calculée d'après g sous échantillons chacun de taille m et $p_{+jg}^{1g} = \sum_i p_{ijg}^{1g}$, $p_{i+g}^{1g} = \sum_j p_{ijg}^{1g}$, que sa valeur augmente avec g de sorte que la probabilité de rejeter l'hypothèse nulle augmente aussi avec g . Hinkins, Oh et Scheuren (1995) ont noté qu'il pourrait être possible de déterminer le nombre de sous échantillons, g , à combiner pour obtenir le seuil de test souhaité en utilisant X^2 . Cette idée semble intéressante, mais l'application effective de la méthode demande une étude plus approfondie, spécialement pour les tests d'hypothèses dans des tableaux multidimensionnels. Au lieu d'utiliser cette approche, il est possible d'élaborer des corrections de Rao Scott de premier et de deuxième ordres pour X^2 en utilisant les sous échantillons multiples pour appliquer les corrections de Rao et Scott (1984) fondées sur le concept des effets de plan. Ces valeurs ajustées seront valides pour tout g . À l'heure actuelle, Benhin étudie les corrections de Rao Scott dans le contexte de l'échantillonnage inverse. À mesure que $g \rightarrow \infty$, le X^2 corrigé convergera vers le X^2 ajusté selon Rao Scott fondé sur l'échantillon complet.

Pour l'échantillonnage PPT sans remise, les praticiens supposent souvent que l'échantillonnage a eu lieu avec remise pour estimer la variance. Dans ce cas, HOS font remarquer qu'il existerait un algorithme de même ordre d'approximation que celui supposé pour estimer les variances (page 16 de HOS).

(v) Nombre de sous échantillons

La stabilité de l'estimateur de la variance d'échantillonage inverse dépend du nombre de sous échantillons, g , tirés à partir de l'échantillon complet et de la fonction (ou du paramètre) qu'on estime. Pour les petites valeurs de g , cet estimateur peut même prendre une valeur négative. En outre, si m est très petit (comme dans le cas de l'échantillonnage double stratifié avec deux grappes par strate), il faut que la valeur de g soit très grande pour obtenir un estimateur de la variance d'échantillonnage inverse stable. Nous pouvons augmenter la valeur de m par les méthodes approximatives mentionnées en (iii) ou en tirant des sous échantillons aléatoires stratifiés, à condition que les exigences relatives à la confidentialité des données ou d'autres considérations n'empêchent pas d'utiliser des sous échantillons stratifiés.

Hinkins note que le nombre de sous échantillons, g , peut être déterminé par "calage" des estimations en échantillonnage inverse et des estimations en échantillon complet, et que la valeur de g résultante pourrait varier considérablement selon le paramètre d'intérêt. Pour illustrer ce dernier point, Hinkins étudie le cas de trois strates et d'une taille minimale d'échantillon de strate égale à deux et calcule le rapport, r , de l'estimateur de la variance d'échantillonnage inverse à l'estimateur de la variance en échantillon complet pour les deux rapports $R_1 = Y_1/X_1$ et $R_2 = Y_2/X_2$. Elle montre que l'utilisation de $g = 1\ 000$ sous échantillons produit un mauvais calage pour R_2 ($r = 0.46$ comparativement à $r = 0.98$ pour $g = 10\ 000$). Ce résultat est un peu surprenant, mais il pourrait être dû à l'instabilité de l'estimateur de la variance d'échantillonnage inverse avec un nombre $m = 2$ de sous échantillons. Hinkins fait remarquer que l'estimateur de la variance d'échantillonnage inverse pour le terme de coordonnée à l'origine B_0 dans le tableau 2 (dénote EBC) pourrait donner des résultats quelque peu erratiques à mesure que la valeur de g augmente. Nous sommes d'accord avec elle, mais il est difficile de résoudre la question de la convergence pour des paramètres non linéaires tels que B_0 . De toute évidence, nous devons poursuivre les travaux sur le choix de la valeur de g pour l'estimation de la variance en échantillonnage inverse. Fritz Scheuren a souligné dans une communication personnelle que l'utilisateur de données sait, cependant, ce que les principaux utilisateurs vont faire, si bien que g peut

Réponse des auteurs

1. INTRODUCTION

Nous remercions les critiques, John Eltinge et Susan Hinkins, d'avoir formulé des commentaires intéressants et proposé certains sujets à explorer plus en profondeur concernant l'échantillonnage inverse. Nous nous efforcerons, dans notre réplique, de répondre à certaines questions qu'ils ont soulevées.

Notre étude de l'échantillonnage inverse a été motivée par les premiers travaux traitant du sujet réalisés par Hinkins, Oh et Scheuren (1997) (appelés ci après HOS). Ces auteurs ont élaboré plusieurs algorithmes d'échantillonnage inverse et offert certaines applications. Ils ont aussi fait remarquer que l'échantillonnage inverse donnait la possibilité de fournir des fichiers de microdonnées à grande diffusion, formés de plusieurs sous échantillons aléatoires simples, utilisables pour faire des inférences valides, comme l'analyse par régression et l'analyse de données nominales, et pour produire des représentations graphiques des données. La contribution principale de notre article est d'offrir un appui théorique (théorèmes 1 à 5) et d'élaborer la méthode des équations d'estimation combinées (EBC) (section 5) qui permet de procéder à diverses analyses des données, comme la régression linéaire et la régression logistique, même si la taille des sous échantillons est faible. Nous élaborons un estimateur par linéarisation de la variance d'échantillonnage inverse (équations (5.17) et (5.20)) qu'on peut calculer à partir du fichier de micro-données et donnons les conditions de sa convergence vers l'estimateur par linéarisation de la variance en échantillon complet à mesure que le nombre de sous échantillons, g , tend vers ∞ .

i) Estimation ponctuelle d'un total

Dans le contexte de l'estimation d'un total $\theta = X$, nous proposons l'estimateur d'échantillonnage inverse \hat{Y}_g donné par (4.1) et montrons que, quand $g \rightarrow \infty$, il converge vers l'estimateur de Horvitz Thompson en échantillon complet sous la condition $\pi_i(s_0) = \pi_i$ pour tout $s_0 = t$ (voir le théorème 3). Eltinge soulève la question importante de l'augmentation de l'efficacité de \hat{Y}_g , pour une valeur de g donnée. À cette fin, il suggère de considérer éventuellement l'échantillonnage inverse simple comme un cas particulier de l'échantillonnage double et, selon cette analogie, d'appliquer des estimateurs en échantillonnage inverse fondés sur la méthode du ratio ou de la régression en construisant un ensemble de données à grande diffusion comprenant g sous échantillons, $\{(y^j, x^j); j \in s_j\}$, $j = 1, \dots, g$, complétés par les totaux estimés en échantillon complet X pour un vecteur de variables auxiliaires, x . Par exemple, un estimateur en échantillonnage inverse par le ratio est donné par $\hat{Y}_{rg} = (Y_g/X_g)X$, où X_g est l'estimateur en échantillonnage inverse du total X . Eltinge fait

ii) Paramètres non linéaires

À la section 3, nous considérons un estimateur en échantillonnage inverse "distinct", $\hat{\theta}_g$, d'un paramètre non linéaire θ , comme un rapport de totaux $\theta = Y/X$, et notons que $\hat{\theta}_g$ peut donner lieu à un biais important si la taille du sous échantillon, m , est faible. Cette situation est due au fait que le biais de $\hat{\theta}_g$ est d'ordre m^{-1} . Dans sa discussion, Hinkins note que HOS étaient, en fait, conscients de ce problème et qu'ils ont commenté brièvement l'estimation du rapport R (page 18 de HOS). En particulier, HOS ont proposé d'estimer séparément le numérateur Y et le dénominateur X , pour aboutir à l'estimateur en échantillonnage inverse "combiné", $R_g = Y_g/X_g$, qui se déduit comme un cas spécial de notre méthode EBC (voir la section 5.2). À la section 5.1, nous exposons la méthode combinée de HOS pour le rapport R , comme nous l'a suggéré le rédacteur adjoint, Fritz Scheuren.

Eltinge note aussi que, dans certains cas, les poids en échantillon complet sont rajustés pour éviter les problèmes de gonflement de la variance causés par les observations influentes. Il se demande s'il est possible de modifier les algorithmes d'échantillonnage inverse de sorte que l'estimateur en échantillonnage inverse résultant, disons \hat{Y}_g , converge vers l'estimateur en échantillon complet à pondération rajustée, disons, \hat{Y} , quand $g \rightarrow \infty$. Ce problème semble difficile, mais on pourrait peut être obtenir des solutions approximatives.

iii) Estimateur de la variance approximative

Eltinge note que des estimateurs approximatifs de la variance en échantillon complet, tels que ceux fondés sur la regroupement des strates, ont été proposés dans la littérature et qu'il pourrait être possible d'élaborer des méthodes d'échantillonnage inverse fondées sur le "plan d'estimation de la variance" approximative plutôt que sur le plan d'échantillonnage original. Ce genre de méthodes pourrait produire des sous échantillons de plus grande taille, m . Par exemple, dans le cas de l'échantillonnage double stratifié avec deux grappes par strate, nous avons $m = 2$ et nous pouvons augmenter la valeur de m en regroupant certaines strates. Ceci, à son tour, pourrait réduire le nombre de sous échantillons, g , nécessaires comparativement à celui requis pour le plan d'échantillonnage

Comme le soulignent correctement RSB dans leurs conclusions, les possibilités d'approfondissement de la recherche et de l'analyse restent nombreuses. Leur article fait progresser considérablement la connaissance théorique et l'application éventuelle des méthodes de rééchantillonnage, ouvrant ainsi d'autres débouchés.

RÉFÉRENCES ADDITIONNELLES

- DOYLE, P., LANE, J., THEEUWES, J. et ZAYATZ, L. (2001). *Confidentiality, Disclosure and Data Access*. North-Holland : New York.
- DUNCAN, G., JABINE, T. et DEWOLF, V. (1993). *Private Lives and Public Policies*. National Academy Press: Washington.
- ELTINGE, J.T. (1999). Use of stratum mixing to reduce primary-unit-level identification risk in public-use survey datasets. *Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference*.
- FELLEGLI, I. (1980). Approximate tests of independence and goodness of fit based on multistage samples. *Journal of the American Statistical Association*, 75, 261-268. Voir aussi Scheuren, F. (1972), Topics in Multivariate Finite Population Sampling and Data Analysis : George Washington University Doctoral Dissertation.
- HINKINS, S., et SCHEUREN, F. (2001). *Increasing Public Accessibility to National Health Interview Survey Data (NHIS) Using Inverse Sampling*. Rapport préparé pour NCHS under a Professional Services Contract.
- HINKINS, S., OH, H.T. et SCHEUREN, F. (1995). Using an inverse sampling algorithm for tests of independence based on stratified samples. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- HINKINS, S., PARSONS, V. et SCHEUREN, F. (2000). Increasing Public Accessibility to Complex Survey Data by Using Inverse Sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- HINKINS, S., LIU, Y. et SCHEUREN, F. (1998). Présentation à l'Annual Statistical Society of Canada Meeting, juin 1998.
- LIU, Y. (1999). Balanced Sampling Design: An Improvement over the Classical Sampling Design. Thèse de doctorat. The George Washington University.
- MULROW, J., et SCHEUREN, F. (1998). The Confidentiality Beasts. *Turning Administrative Systems into Information Systems*. Internal Revenue Service.
- PARSONS, V.T., et ELTINGE, J.T. (1999). Stratum partition, collapse and mixing in construction of balanced repeated replication variance estimators. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- des totaux annuels, on utilise un plan d'échantillonnage fortement stratifié. Cependant, pour les économistes, la modélisation de l'activité économique et l'élaboration de modèles fiscaux, qui ne posent pas le même problème que le calcul d'estimations par régression en population finie, sont d'autres utilisations importantes des données.
- Un autre exemple de ce genre, tiré de l'EPA, est un grand échantillon stratifié des lacs américains sur l'eau desquels des analyses chimiques ont été faites afin de fournir des données de base relatives aux pluies acides. Ces données intéressaient aussi beaucoup les biologistes qui souhaitaient modéliser certains aspects des relations chimiques et physiques.
- L'interprétation des modèles de régression dans le cas de l'échantillonnage en population finie peut prêter à confusion. Il existe de nombreuses méthodes bien pensées de régression dans des conditions d'échantillonnage complexes, mais la règle empirique simplifiée est qu'on ne peut généralement pas ignorer la structure du plan d'échantillonnage (par exemple, les poids de sondage). Les analystes qui s'intéressent à la modélisation de la structure paramétrique sous-jacente pourraient considérer ceci comme étant contraire à l'intuition. En outre, si l'on ne tient pas compte du plan de sondage, on peut obtenir une réponse incorrecte, à moins qu'aucune variable explicative ne manque ou que le plan de sondage ne se confonde pas avec les variables explicatives (deux situations peu probable dans le cas des plans complexes). Un échantillon aléatoire simple satisfait la deuxième exigence. Dans le cas de l'échantillon SR, si les économistes souhaitent modéliser la structure des petites et moyennes entreprises, par exemple, des échantillons aléatoires simples assez grands pourraient être générés à partir du plan de sondage stratifié. En outre, la combinaison de tirages multiples pourrait produire une base de données raisonnable.
- Enfin, il est très important d'utiliser des méthodes graphiques en analyse par modélisation et par régression afin de comprendre comment une variable dépend des autres variables explicatives. Même dans le cas simple d'un ou de deux prédicteurs d'une variable indépendante, la représentation graphique au moyen de données d'échantillon pondérées est difficile. L'analyse des résidus et le dépistage des valeurs aberrantes sont plus difficiles quand les données sont pondérées. La représentation graphique, qui est un outil puissant d'extraction d'information à partir des données, semble être un domaine où l'utilisation de l'échantillonnage inverse dans le but de produire des échantillons aléatoires simples devrait être envisagée.

Ainsi, aux Etats Unis, la National Health Interview Survey (NHIS) comprend une stratification au niveau de l'Etat et la sélection de comités et de régions métropolitaines pour l'échantillon. Un fichier de données à grande diffusion de la NHIS est produit sous une forme où la structure complexe de l'échantillon est simplifiée de sorte qu'un plan de sondage stratifié à 2 UPE soit intégré dans chaque strate. Les strates et les UPE du plan de sondage original sont masqués en partie grâce à l'utilisation de certaines techniques discutées dans Eltinge (1999) et dans Parsons et Eltinge (1999). Ce plan masqué à "2 UPE par strate" peut être utilisé pour calculer les variances. Nous avons étudié le plan de sondage de la NHIS pour voir si l'on pouvait appliquer l'échantillonnage inverse pour la production de données à grande diffusion (Hinkins et Scheuren 2001) et constaté qu'il était impossible d'inverser le plan de sondage jusqu'au niveau de détail utile pour les analystes des données. Nous restons convaincus que l'échantillonnage inverse peut être une option séduisante pour la production d'ensembles de données à grande diffusion lorsque le plan de sondage est inversible. Toutefois, il ne s'agit pas nécessairement d'une option viable, à moins qu'on ait prévu son utilisation au stade de l'élaboration du plan de sondage original, de façon à utiliser des plans de sondage inversibles.

Une autre utilisation possible de l'échantillonnage inverse mérite d'être mentionnée en ce qui concerne le présent exemple. Pour les domaines analytiques couvrant la plupart des strates, les estimateurs de la variance calculés d'après les données à grande diffusion de la NHIS sont stables, autrement dit le nombre de degrés de liberté associés aux estimateurs est grand. Par contre, pour les sous populations moins dispersées géographiquement, qui couvrent un moins grand nombre de strates, le nombre de degrés de liberté résultant peut être assez faible et les estimations de la variance, assez instables. Le cas échéant, il serait peut être possible de produire un estimateur plus stable de la variance en tirant un très très grand nombre d'échantillons d'après le plan de sondage pour données à grande diffusion. Dans ce cas, au lieu de fournir des données à grande diffusion, on pourrait utiliser l'échantillonnage inverse comme calculateur "à boîte noire" de la variance qui donnerait des estimateurs plus stables de la variance pour les éléments rares de la population.

Applications de modélisation et de représentation graphique – On peut utiliser l'échantillonnage inverse pour produire des données sous une forme offrant plus de possibilités d'analyse. Cette option serait particulièrement intéressante lorsque l'usage prévu des données est multiple. Un exemple naturel se dégage de notre proposition initiale d'utiliser une méthode de rééchantillonnage pour les échantillons stratifiés de déclarations des bénéficiaires des sociétés pour la statistique du revenu (SR). La population sous jacente étant fortement asymétrique (un nombre assez faible de grandes unités représente une forte proportion de la valeur totale), afin de produire des estimations efficaces méthodes standard.

La suite de la discussion porte sur deux domaines d'intérêt où l'échantillonnage inverse pourrait être utile, à savoir la production de données à grande diffusion et l'analyse par modélisation ou par régression. Ces deux problèmes illustrent aussi deux types généraux d'utilisation inverse pouvant avoir pour objectif de produire des données à grande diffusion qui donneront de meilleures estimations en grande partie similaires aux estimations obtenues d'après le plan de sondage complexe, tout en permettant d'appliquer les méthodes d'analyses courantes à l'inférence fondée sur des méthodes d'échantillonnage inverse concorde avec celle fondée sur des méthodes d'inférence fondée sur des méthodes d'échantillonnage inverse complexe complet, les utilisateurs de données dont les ressources informatiques sont limitées pourront exécuter certaines analyses fondées sur le plan de sondage au moyen de logiciels statistiques courants. Les résultats présentés dans l'article de RSB étendent la théorie et offrent des conditions où l'utilisation de ces techniques de rééchantillonnage sont applicables.

Une caractéristique nécessaire des données à grande diffusion est la protection des renseignements confidentiels. Pour les organismes statistiques fédéraux des Etats Unis, les fichiers de données à grande diffusion ont été un des moyens de réaliser l'objectif de "transparence" (par exemple, Duncan et coll. 1993). Cependant, la diffusion de données fréquentes de données électoniques de toutes sortes par la voie d'Internet et les progrès dans le domaine des logiciels de couplage d'enregistrements peuvent être considérés comme des menaces pour cette transparence (par exemple, Doyle et coll. 2001).

Les objectifs relatifs aux données à grande diffusion peuvent devenir contradictoires, par exemple, s'il faut fournir, implicitement ou explicitement, des renseignements sur la méthode d'échantillonnage pour permettre le calcul de la variance par rapport au plan de sondage, mais que ces renseignements augmentent considérablement la probabilité d'identifier un individu. Dans de nombreuses enquêtes, l'emplacement géographique joue un rôle important dans l'échantillonnage, mais on ne peut diffuser les détails les plus fins de la structure géographique de l'échantillonnage en même temps que les données sans compromettre le respect de la confidentialité des renseignements personnels fournis par les individus. Si l'on supprime les renseignements sur la structure géographique d'échantillonnage pour assurer la confidentialité, les données deviennent plus difficiles à analyser par les méthodes types fondées sur le plan de sondage. Le cas échéant, l'utilisation de l'échantillonnage inverse permettrait de diffuser des données sans détails les plus fins de la structure géographique, par exemple, tout en permettant encore l'analyse par des méthodes standard.

originales; en utilisant 10 000 répétitions, elles sont situées dans un intervalle de $\pm 0,3\%$ par rapport à ces estimations. Par contre, les estimations des erreurs types ont un comportement assez différent. Pour chaque paramètre, le tableau 1 montre le ratio de l'erreur type estimée pour les échantillons aléatoires simples combinés à l'erreur type estimée de l'estimation sur l'échantillon stratifié original. Nous avons estimé la variance des estimations en échantillon combiné selon la méthode décrite plus bas.

Tableau 1

Ratio des erreurs types estimées : estimation combinée par rapport à l'estimation originale

Paramètre	1 000 échantillons	10 000 échantillons
Totaux	1.22	1.03
X_1	1.21	0.99
Y_1	1.07	1.07
Estimation du $R_1 = Y_1/X_1$	1.07	
Totaux	1.02	0.95
X_2	0.94	0.93
Y_2	0.46	0.98
Estimation du $R_2 = Y_2/X_2$	0.46	
ratio		

Si l'on utilise 1 000 échantillons aléatoires simples d'après l'échantillon combiné est raisonnablement proche de celle de l'estimation d'après l'échantillon original. Entre parenthèses, ce résultat ne nous étonne pas. Si l'on utilise 10 000 répétitions, l'erreur type de l'estimation estimée de l'estimation de X_1 d'après l'échantillon stratifié d'après l'échantillon combiné est supérieure de 22 % à l'erreur type de l'estimation de X_1 d'après l'échantillon combiné. Cette constatation est relativement stable.

Considérons le deuxième ensemble de variables. Cette fois-ci, les erreurs types des estimations des totaux X_2 et Y_2 d'après l'échantillon combiné semblent converger en utilisant seulement 1 000 échantillons. Par contre, l'erreur type de l'estimation de R_2 est fortement sous estimée, comparativement à l'erreur type de l'estimation d'après l'échantillon original stratifié. Toutefois, 9 000 tirages supplémentaires augmentent l'erreur type estimée pour le ratio au point qu'elle est approximativement égale à celle de l'estimateur original.

De toute évidence, il est nécessaire de poursuivre l'analyse de l'utilisation de l'échantillonnage inverse et de l'estimation de la variance pour les estimations par le ratio et par la régression. En outre, le présent exemple souligne que l'échantillonnage inverse doit tenir compte de tous les paramètres d'intérêt.

Comme on pourrait s'y attendre, l'estimateur de la variance pour l'estimation par le ratio combinée est le même que celui construit par Rao, Scott et Benhin par leur méthode des équations d'estimation.

L'utilisation des répétitions combinées de l'échantillon pour estimer les coefficients de régression est également examinée par RSB. Ils ont élaboré une estimation de la variance, par la méthode des équations d'estimation, qui donne de bons résultats et qui étend les possibilités d'utilisation des techniques de rééchantillonnage. Leurs résultats permettent aussi de poursuivre l'étude des propriétés de la variance estimée pour des échantillons combinés.

Dans l'exemple de régression de RSB, il n'est pas certain que l'estimation de la variance pour B_0 ait convergé. La question de la convergence des estimations de l'erreur des paramètres non linéaires est intéressante, étant donné surtout qu'il est probable qu'on utilise ces estimations pour étalonner le processus. (Par échantillonnage, nous entendons la détermination du moment où un nombre "suffisant" d'échantillons ont été tirés, compte tenu de l'usage souhaité). Dans le cas de l'estimation d'un paramètre, la seule information disponible pour l'échantillonnage pourrait être la comparaison de l'estimation combinée, par exemple, à l'estimation originale fondée sur le plan d'échantillon complexe, ainsi que la comparaison de leurs erreurs types estimées. Autrement dit, nous savons peut être que la variance convergera, mais nous ne disposons que des estimations de la variance pour l'échantillonnage.

Considérons l'exemple qui suit où nous utilisons l'algorithme d'échantillonnage inverse pour inverser un plan de sondage comportant trois strates pour lequel la taille minimale de strate est égale à deux. Par conséquent, chaque répétition est de taille $m = 2$ et on ne devrait pas s'attendre à une convergence rapide. Nous estimons les deux ratios. En utilisant 1 000 répétitions, les estimations ponctuelles calculées à partir des échantillons combinés sont comprises dans un intervalle de $\pm 1,0\%$ par rapport aux estimations

$$\text{var}(R_c) = \frac{1}{1} \sum_{j=1}^J \left(\frac{1}{m} - \frac{1}{N} \right) s_{f_e}^2 - \frac{1}{1} \sum_{j=1}^J \left(\frac{1}{m} - \frac{1}{N} \right) s_{f_e}^2$$

$$\text{où } s_{f_e}^2 = \frac{1}{m} \sum_{i=1}^m (e_{fi} - \bar{e}_f)^2$$

où $\bar{e}_f = (1/g) \sum_{j=1}^J e_{fj}$ et la moyenne du J^{e} échantillon répété est $\bar{e}_f = \bar{x}_j - R_c \bar{y}_j = \bar{x}_j - (\bar{y}_j / \bar{X}_j) \bar{Y}_j$.

L'utilisation de l'estimation de la variance généralisée par l'équation (3.4) de RSB pour estimer la variance de e_c donne l'estimation de la variance suivante pour l'estimation par le ratio combinée :

$$\text{var}(R_c) = \text{var} \left(\frac{Y_c}{X_c} \right) = \frac{1}{1} \text{Var}(e_c)$$

Commentaires

SUSAN HINKINS¹

consiste à déterminer la façon d'élaborer les tests, afin d'obtenir le niveau souhaité de signification (par exemple, un niveau de signification de 0,05). Les premiers résultats indiquent qu'on pourrait déterminer le nombre d'échantillons aléatoires simples à combiner pour obtenir le niveau de signification souhaité pour le test et que l'utilisation du chi-carré de Pearson sur les échantillons combinés donne des résultats qui se comparent favorablement à ceux de l'application de la méthode de Fellegi (1980) à l'échantillon stratifié original, en étant peut-être plus facile à utiliser.

À la conférence de 1998 de la Société statistique du Canada, Hinkins, Liu et Schreuen ont présenté des résultats de simulation pour l'ajustement de modèles de régression à des échantillons inverses obtenus à partir d'un plan de sondage complexe (plan stratifié équilibré par rapport à la médiane). Dans ce cas, le plan original prévoyait la

sélection de 100 répliques; dans chaque réplique, une observation a été sélectionnée à partir de chacune des six strates, de sorte que les observations soient équilibrées par rapport à la médiane (Liu 1999). La sélection a été faite avec remise dans toutes les répliques. L'échantillonnage inverse consistait à sélectionner une unité à partir de chaque réplique. Nous avons examiné l'ajustement d'un modèle de régression aux échantillons inverses individuels et l'ajustement d'un modèle de régression à la combinaison de plusieurs échantillons inverses. Pour la droite de régression les échantillons inverses uniques, la pente estimée variait de 0,70 à 1,13. Pour les six échantillons inverses combinés, la pente estimée était de 0,845 avec $R^2 = 0,64$.

Estimation de la variance – L'estimation de la variance selon la formule donnée par HOS en 1997 pose un problème intéressant, car les échantillons ne sont pas inconditionnellement indépendants. Dans notre article de 1997, nous avons suggéré que, dans le cas de l'estimation par le ratio, si l'échantillon combiné est suffisamment grand pour que l'approximation par développement en série de Taylor soit acceptable, il serait peut-être possible d'utiliser l'approximation "habituelle" de la variance pour un ratio. Autrement dit, on pourrait estimer la variance en utilisant

$$\text{Var}(\hat{R}) = \frac{1}{L} \text{Var}(\bar{e}) \text{ où } \bar{R} = \frac{X}{Y} \text{ et } e_i = y_i - R x_i.$$

La variance estimée de l'estimation par le ratio fondée sur les échantillons combinés peut alors être calculée de la façon "habituelle".

Rao, Scott, et Benhin (RSB) ont fort bien résumé nos résultats sur les échantillons complexes inversibles et ont fait avancer considérablement l'étude du sujet grâce à un ensemble impressionnant de résultats théoriques. Leur article offre de nouvelles perspectives précieuses pour les statisticiens qui souhaitent considérer, à l'étape de la conception du plan de sondage, l'option d'utiliser des méthodes de rééchantillonnage durant l'analyse. De cette façon, il est possible d'utiliser des plans de sondage inversibles. Comme le font remarquer les auteurs, nombre de problèmes intéressants restent encore à résoudre. Nous discutons ici de certains points particuliers de l'article, ainsi que de certaines questions débatables soulevées durant nos travaux de recherche appliquée sur l'emploi de l'échantillonnage inverse.

Façon d'utiliser les échantillons résultants – L'estimation de totaux ou de moyennes présente l'avantage de donner des estimations identiques d'après des échantillons combinés ou distincts. Au-delà des problèmes d'estimation (simple), nombre de questions débatables se posent quant à la meilleure utilisation des échantillons résultants, mais la combinaison de ces derniers est une option fort raisonnable. Pour un paramètre tel qu'un ratio, c'est à dire une fonction de totaux, il semble intuitif de calculer la meilleure estimation de chaque total, puis d'appliquer la fonction aux estimations, et cette démarche est celle que nous recommandons. En fait, comme l'estimateur par le ratio est utilisé dans nombre de situations, nous avons commenté brièvement cette question dans l'article publié par HOS en 1997. Toutefois, RSB ont prouvé ce point explicitement et, de surcroît, ont donné une méthode cohérente d'estimation de la variance à partir d'échantillons combinés. Ceci fournit aux chercheurs des outils utiles pour l'application des techniques d'échantillonnage inverse à une gamme plus variée de problèmes.

Dans Hinkins, Oh et Schreuen (1995), nous considérons la situation de l'échantillonnage inverse pour le calcul de tests d'indépendance à partir d'un tableau de contingence 2x2 lorsque les données proviennent d'un échantillon stratifié. L'analyse par tableau de contingence et l'analyse par régression ont toutes deux été mises au point en grande partie dans l'univers III et, par conséquent, des corrections sont nécessaires lorsqu'on les applique à des données provenant d'enquêtes complexes. Nous avons tiré plusieurs échantillons aléatoires simples et calculé la simple variable chi-carré de Pearson à partir des données combinées. À mesure que le nombre d'échantillons augmente, la probabilité de rejeter l'hypothèse nulle augmente aussi, si bien qu'on ne peut pas prendre un nombre arbitrairement large d'échantillons aléatoires simples. Le problème

RUST, K., et KALTON, G. (1987). Strategies for collapsing strata for

variance estimation. *Journal of Official Statistics*, 3, 69-81.

SÄRNDAAL, C.-E., SWENSSON, B. Et WREFTMAN, J. (1992).

Model-Assisted Survey Sampling. New York : Springer-Verlag.

SCHUREN, F.J. (1997). Communication personnelle.

DE WAAL, A.G., et WILLENBORG, L.C.R.J. (1997). Statistical

disclosure control and sampling weights. *Journal of Official*

Statistics, 13, 417-434.

ZASLAVSKY, A.M., SCHENKER, N. et BELIN, T.R. (2001).

Downweighting influential cliques in surveys: Application to the

1990 Post Enumeration Survey. *Journal of the American*

Statistical Association, 96, 858-869.

RÉFÉRENCES ADDITIONNELLES

CHEN, G., et KELLER-MCNULTY, S. (1998). Estimation of

identification disclosure risk in microdata. *Journal of Official*

Statistics, 14, 79-95.

DUNCAN, G.T., et PEARSON, R.W. (1991). Enhancing access to

microdata while protecting confidentiality (avec discussion).

Statistical Science, 6, 219-239.

KORN, E.J., et GRAUBARD, B.I. (1995). Analysis of large health

surveys: Accounting for the sampling design. *Journal of the*

Royal Statistical Society, Series A 158, 263-295.

MANTEL, H. (2002). Discussion à Statistique Canada Symposium.

8 Novembre 2002.

5. REPRÉSENTATIONS GRAPHIQUES

Hinkins et coll. (1997, page 19) et Scheuren (1997) ont souligné les possibilités d'application de l'échantillonnage inverse à la représentation statistique graphique des données d'enquête complexe. Par exemple, Scheuren (1997) note que de nombreuses méthodes statistiques graphiques (par exemple, les nuages de points) ont été mises au point principalement pour des ensembles d'observations indépendantes et identiquement distribuées. Les applications directes de ces méthodes aux données d'enquête complexe peuvent produire des graphiques d'indicateurs d'unité primaire d'échantillonnage nominal-graphiques et au niveau du ménage X peut donner lieu à l'identification d'une unité primaire d'échantillonnage si les agrégats des observations X au niveau de l'UPB varient selon des profils distincts connus du public. Par exemple, un comté donné pourrait avoir un profil démographique inhabituel ou un profil de dépenses distinct, par exemple, pour le gaz naturel ou l'électricité.

Pour cette raison, il serait intéressant d'évaluer la mesure dans laquelle la grande diffusion de données obtenues auprès d'échantillons inverses multiples pourrait fournir des renseignements permettant à un utilisateur de données de reconstruire les poids no net des groupements au niveau de l'UPB qui sont informatifs. Par exemple, en nous inspirant des commentaires formulés par Mantel (2002), supposons que la mesure d'une variable donnée X soit déclarée sur une échelle continue et que, pour nombre d'unités répondantes, la valeur numérique de X soit unique, alors (conformément aux commentaires de la section 2) l'appariement des valeurs déclarées de X pour un très grand nombre g d'échantillons inverses multiples permettrait à un utilisateur des données d'estimer les poids probabilistes associés à un répondant particulier i . Ceci, à son tour, pourrait nous ramener aux problèmes d'identification susmentionnés considérés par de Waal et Willemborg (1997) et par Chen et Keller-McNulty (1998). Dans certains cas extrêmes, des problèmes d'identification similaires pourraient survenir pour les unités primaires d'échantillonnage. L'importance pratique de ces préoccupations dépend de la grandeur empirique relative des diverses sources d'erreur (y compris l'erreur induite par l'utilisation d'un nombre fini d'échantillons g) et pourrait représenter un sujet d'étude intéressant pour certains cas particuliers à un organisme.

6. RISQUE D'IDENTIFICATION

L'auteur remercie Van Parsons, Fritz Scheuren et Al Zarate pour les nombreuses discussions fructueuses de l'échantillonnage inverse et de ses utilisations possibles en vue de réduire le risque d'identification. Les opinions exprimées ici sont celles de l'auteur et ne reflètent pas forcément les politiques du U.S. Bureau of Labor Statistics.

REMERCIEMENTS

4. SIMPLICITÉ OPÉRATIONNELLE

En principe, la plupart des estimations ponctuelles, des estimations de la variance et des méthodes d'inférence qui

ont été élaborées pour les données obtenues par échantillonnage aléatoire simple peuvent être étendues aux données obtenues par échantillonnage complexe. Cependant, les efforts qu'exige ce genre d'extension ne sont pas négligeables et pourraient dissuader nombre d'analyses éventuelles et les analyses qui, selon eux, offriront le plus d'éclaircissements grossiers de rentabilité, où ils se concentrent sur les leurs méthodes analytiques en se basant sur une évaluation informel, les analystes de données semblent souvent choisir d'utiliser efficacement les données disponibles. En un sens, qu'ils considèrent disproportionné par rapport aux avantages scientifiques éventuels. Souvent, les statisticiens et les analystes des domaines spécialisés ont des points de vue différents en ce qui concerne les coûts et les avantages scientifiques relatifs d'un effort analytique particulier. Dans certains cas, l'échantillonnage inverse peut contribuer à amoindrir les effets de ces opinions divergentes.

En particulier, comme l'indiquent RSB et HOS, l'investissement d'un organisme statistique dans la construction d'échantillons inverses peut donner lieu à une réduction de la charge de travail d'un analyste donné. Cet investissement peut être particulièrement précieux si les deux conditions qui suivent sont satisfaites.

a) Un analyste a l'intention d'exécuter un grand nombre d'analyses différentes sur un ensemble de données d'enquête unique, mais il ne possède pas le logiciel approprié pour les enquêtes complexes pour nombre des analyses prévues, voire toutes, et perçoit la programmation de procédures applicables aux enquêtes complexes comme un effort important.

b) Les étapes supplémentaires de calcul nécessaires pour l'estimation ponctuelle (par exemple, le calcul de la moyenne effectué dans les estimateurs ponctuels (3.1) ou (4.1), ou les équations d'estimation combinées (5.14) ou l'estimation de la variance (par exemple, les estimateurs de la variance (3.3), (3.4), (5.18) ou (5.20)) imposent une charge de travail supplémentaire relativement faible à l'analyste ou peuvent être absorbées dans le logiciel analytique de façon transparente pour l'analyste.

naturelle est celle de savoir si l'on pourrait modifier l'algorithme d'échantillonnage inverse de sorte que le plan de sondage inverse soit "régler" sur les poids corrigés plutôt que sur les poids probabilistes inverses directs. Cette option serait fort intéressante pour les cas où l'on s'attend à ce que l'erreur quadratique moyenne des estimateurs ponctuels à pondération corrigée soit nettement plus faible que celle des estimateurs ponctuels pondérés par probabilité inverse. Pour les cas où cette méthode modifiée est souhaitable, il serait intéressant d'étudier les moyens correspondant d'étendre l'approche de RSB à l'estimation de la variance.

3. APPROXIMATIONS UTILISÉES POUR L'ESTIMATION DE LA VARIANCE ET L'INFÉRENCE

Pour certains plans de sondage complexes, HOS et RSB ont noté que l'extraction exacte d'un échantillon aléatoire simple peut être impossible ou donner lieu à un échantillon inverse très petit qui, à son tour, doit être compensé par l'utilisation d'une valeur très grande de g . Par conséquent, dans les sections 2 et 4.3 de leur article, RSB considèrent des méthodes d'approximation inverse et, à la section 4.1, ils considèrent l'échantillonnage inverse qui pourrait produire un plan de sondage plus simple que le plan complexe original, mais plus compliqué que le plan d'échantillonnage aléatoire simple.

Parallèlement, rappelés que certaines méthodes d'échantillonnage décrites dans la littérature envisagent des estimateurs de la variance fondés sur des approximations du plan d'échantillonnage réel. Un exemple est l'estimation de la variance fondée sur le regroupement des strates. Voir, par exemple, Rust et Kalton (1987) et les références citées par ces auteurs. En outre, Korn et Graubard (1995, sections 4.2 et 4.3) considèrent des estimateurs de la variance qui ne tiennent pas compte de la mise en grappes originale au niveau des unités primaires d'échantillonnage et traitent les unités secondaires d'échantillonnage comme si elles étaient des unités primaires d'échantillonnage.

Dans certains cas, ces approches peuvent être problématiques, tandis que dans d'autres, elles peuvent produire des estimateurs de la variance satisfaisants. Dans ces derniers cas, on pourrait considérer l'élaboration d'une méthode d'échantillonnage inverse fondée sur le "plan d'estimation de la variance" approximative plutôt que sur le plan de sondage réel. Sous cette approche, il serait particulièrement intéressant de considérer la grandeur relative des erreurs associées à, respectivement, l'échantillonnage sous le plan original, l'erreur d'approximation dans le "plan d'estimation de la variance" et l'erreur supplémentaire induite par l'utilisation d'un nombre fini d'échantillons inverses.

1. VUE D'ENSEMBLE

ponds dans l'estimateur ponctuel d'Horvitz-Thompson à mesurer que λ augmente. En ce sens, nous pouvons considérer l'estimateur ponctuel (4.1) comme une approximation de l'estimateur d'Horvitz-Thompson. Des commentaires semblables s'appliquent aux estimateurs ponctuels non linéaires généraux et aux échantillons inverses généraux considérés dans RSB.

En outre, l'échantillonnage inverse simple peut être considéré comme une forme spéciale d'échantillonnage double dans lequel les taux de sélection à la deuxième phase sont proportionnels à l'inverse des taux d'échantillonnage à la première phase. Ceci nous mène naturellement à la question de savoir si les idées classiques se dégagent de l'échantillonnage double peuvent aboutir à des gains d'efficacité en échantillonnage inverse simple multiples.

Par exemple, rappelons que, dans l'échantillonnage double classique, on peut souvent améliorer l'efficacité en utilisant conjuguées à des variables auxiliaires X observées pour toutes les unités d'échantillonnage de première phase. Consulter, par exemple, Särndal, Swenson et Wretman (1992, chapitre 9). Pareillement ici, on pourrait construire un ensemble de données à grande diffusion correspondant à un ensemble de données obtenues sur échantillon inversé

simple ou multiple accompagnée des totaux estimés (calculés en se fondant sur l'échantillon complexe complet) pour un vecteur de variables supplémentaires X . En outre, certaines données auxiliaires supplémentaires pourraient être nécessaires pour obtenir une estimation de la variance convergente. À condition que les variables auxiliaires X soient suffisamment fortes, les estimateurs ponctuels corrigés par le ratio ou par la régression pourraient contribuer à l'amélioration de la précision des analyses fondées sur l'échantillonnage inverse. Ceci pourrait, à son tour, réduire le nombre de sous-échantillons inverses nécessaires pour assurer que la variance de l'estimateur ponctuel en échantillonnage inverse multiple corrigé par régression soit suffisamment faible.

De façon plus générale, dans de nombreux cas d'enquête complexe (n'appartenant pas à la catégorie des plans de sondage double), les estimateurs pondérés pondérés proviennent aussi au delà de l'utilisation directe de poids probabilistes inverses pour intégrer l'information auxiliaire grâce à, par exemple, des corrections par ratio ou par régression. De plus, dans certains cas, on réduit les valeurs numériques de certains poids probabilistes extrêmes pour essayer d'éviter les problèmes de gonflement de la variance causés par les observations influentes. Voir, par exemple, Zaslavsky, Schenker et Belin (2001). Une question

Rao, Scott et Benhin (ci après RSB), ainsi que Hinkins, Oh et Scheuren (1997) (ci après HOS), ont produit un

L'échantillonnage inverse. La présente discussion nous soulera un ensemble fascinant d'idées et de méthodes ayant trait à l'échantillonnage inverse. La présente discussion nous soulera plusieurs idées et questions pratiques connexes que devront vraisemblablement examiner les statisticiens d'enquête qui envisagent certaines applications pratiques de l'échantillonnage inverse. La section 2 souligne certaines relations entre la pondération probabiliste type et la pondération inverse répétée. La section 3 décrit deux types d'approximation de l'échantillonnage inverse intégrée dans l'échantillonnage aléatoire implicite. La section 4 décrit les simplifications pratiques qui pourraient résulter de l'utilisation de l'échantillonnage inverse dans certains cas. La section 5 porte sur l'utilisation des données d'échantillonnage inverse dans les méthodes géographiques (fondée sur un échantillon aléatoire simple). La section 6 explore les avantages et les limites éventuelles de l'échantillonnage inverse lors d'efforts en vue de réduire le risque d'identification dans les ensembles de données à grande diffusion.

2. ESTIMATION PONCTUELLE :

En empruntant certaines idées publiées sur l'échantillonnage, le traitement de signaux et la confidentialité (par exemple, Duncan et Pearson, 1991), nous pouvons considérer un estimateur ponctuel comme étant le résultat de plusieurs étapes de "filtration" d'observations faites sur une population. Par exemple, dans la construction d'un estimateur type de Horvitz-Thompson d'un total de population, un ensemble de valeurs de population peuvent être considérées comme passant par deux étapes de filtration correcte-pondant, respectivement, à la sélection des unités d'échantillonnage et à la pondération de ces unités par la probabilité inverse de sélection. Parallellement, l'estimateur ponctuel (4.1) dans RSB peut être considéré comme le résultat de deux étapes de filtration, où la deuxième correspond maintenant à la pondération par un facteur aléatoire déterminé d'après le nombre de fois qu'une unité d'échantillonnage donne figure dans les g échantillons inverses répétés. Dans ces conditions, les poids filtrés dans (4.1) convergent vers les probabilités inverses utilisées comme

Preuve du théorème 3

Nous avons

$$Y_i^* = \sum_{i \in s_0^*} \frac{Y_i^* \pi_i^*}{Y_i^* \pi_i^*} = \sum_{i \in s_0^*} \frac{Y_i^* \pi_i^*}{Y_i^* \pi_i^*},$$

où $Y_i^*(s_0^*)$ prend la valeur 1 si la i^e unité est incluse dans le j^e sous-échantillon s_j^* et 0 autrement, et π_i^* est la probabilité d'inclusion (inconditionnelle) correspondante.

Donc

$$Y_i^\infty = E[Y_i^* | s_0] = \sum_{i \in s_0} \frac{\pi_i^*}{Y_i^* \pi_i^*}.$$

Cette expression est égale à $Y_i = \sum_{i \in s_0} (Y_i / \pi_i)$, c'est-à-dire l'estimateur H-T pour le plan de sondage original si, et uniquement si, $\pi_i(s_0) = \pi_i^* / \pi_i$.

Preuve du théorème 4

Conditionnellement à s_0 , il découle de (3.3) que $Y_{i,HT}^g$ converge presque certainement vers

$$Y_{i,HT}^\infty = E[Y_i^* | s_0] - \text{Var}(Y_i^* | s_0) \quad (\text{A.1})$$

quand $g \rightarrow \infty$. Maintenant, en notant que $\pi_{ii}^*(s_0) = \pi_{ii}^* / \pi_i^*$, nous obtenons

$$E[Y_{i,HT}^g | s_0] = \sum_{i \in s_0} \frac{\pi_{ii}^*}{\pi_i^* - \pi_i^*} \frac{\pi_i^* \pi_i^*}{\pi_i^* Y_i^*} = \sum_{i \in s_0} \frac{\pi_{ii}^*}{\pi_i^*} \frac{\pi_i^*}{Y_i^*}.$$

$$= \sum_{i \in s_0} \left(\frac{\pi_{ii}^*}{\pi_i^*} - \frac{1}{Y_i^*} \right) \pi_i^* = \sum_{i \in s_0} \left(\frac{\pi_{ii}^*}{\pi_i^*} - \frac{1}{Y_i^*} \right) \pi_i^* \pi_i^*.$$

(A.2)

En outre,

$$\text{Var}(Y_i^* | s_0) = \sum_{i \in s_0} \left(\pi_{ii}^* - \pi_i^* \pi_i^* \right) \frac{Y_i^* \pi_i^*}{Y_i^* \pi_i^*} = \sum_{i \in s_0} \left(\pi_{ii}^* - \pi_i^* \pi_i^* \right) \frac{Y_i^* \pi_i^*}{Y_i^* \pi_i^*}.$$

(A.3)

Il découle maintenant de (A.1) - (A.3) que $Y_{i,HT}^\infty = Y_{i,HT}^g$.

Preuve du théorème 5

Conditionnellement à s_0 , il découle de (3.3) que

$$Y_{i,SYG}^\infty = E[Y_{i,SYG}^* | s_0] - \text{Var}(Y_{i,SYG}^* | s_0) \quad (\text{A.4})$$

où

$$\text{Var}(Y_{i,SYG}^* | s_0) = \sum_{i < l \in s_0} \left(\pi_i^* \pi_l^* - \pi_{il}^* \right) \left(\frac{Y_i^* \pi_i^*}{Y_l^* \pi_l^*} - \frac{Y_l^* \pi_l^*}{Y_i^* \pi_i^*} \right). \quad (\text{A.5})$$

à condition que la taille des sous-échantillons soit fixe (Cochran 1977, page 260). En outre,

Association, 83, 28-36.

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

BINDER, D.A. (1992). Fitting Cox's proportional hazard models from survey data. *Biometrika*, 79, 139-147.

COCHRAN, W.G. (1977). *Sampling Techniques*. Troisième édition; New York : John Wiley & Sons, Inc.

HINKINS, S., OH, H.T. et SCHEUREN, F. (1997). Algorithmes de plan de sondage inverses. *Techniques d'enquête*, 23, 13-24.

HOFFMAN, E.B., SEN, P.K. and WEINBERG, C.R. (2001). Within-cluster resampling. *Biometrika*, 88, 1121-34.

KOVACEVIC, M.S. et BINDER, D.A. (1997). Variance estimation for measures of income inequality and polarization - the estimating equations approach. *Journal of Official Statistics*, 13, 41-58.

RAO, J.N.K., et SCOTT, A.J. (1992). A simple method for the analysis of clustered binary data. *Biometrics*, 48, 577-585.

RAO, J.N.K., et SCOTT, A.J. (1999). A simple method for analyzing overdispersion in clustered Poisson data. *Statistics in Medicine*, 18, 1373-1385.

SKINNER, C.J., HOLT, D. et SMITH, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. Chichester : Wiley.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York : Springer-Verlag.

BIBLIOGRAPHIE

En comparant (A.7) et (A.4), nous voyons que

$$Y_{i,SYG}^\infty = \sum_{i < l \in s_0} \left(\frac{\pi_{il}^*}{\pi_i^* \pi_l^* - \pi_{il}^*} \right) \left(\frac{Y_i^* \pi_i^*}{Y_l^* \pi_l^*} - \frac{Y_l^* \pi_l^*}{Y_i^* \pi_i^*} \right). \quad (\text{A.7})$$

Il s'ensuit maintenant que

$$E(Y_{i,SYG}^* | s_0) = \sum_{i < l \in s_0} \left(\frac{\pi_{il}^*}{\pi_i^* \pi_l^* - \pi_{il}^*} \right) \left(\frac{Y_i^* \pi_i^*}{Y_l^* \pi_l^*} - \frac{Y_l^* \pi_l^*}{Y_i^* \pi_i^*} \right). \quad (\text{A.6})$$

Tableau 1									
Estimation du ratio de population R									
g = 500									
g = 1 000									
g = 5 000									
Echantillon complet									
EEC									
EES									
Est.	0,4096	0,4101	0,4100	0,4096	0,4095	0,4095	0,4095	0,4094	0,4094
Est. de la var. $\times 10^{-4}$	1,9513	1,8769	1,8508	1,8482	1,8302	1,9320	1,9178		

Tableau 2 Estimation des paramètres de régression « en cas de recensement », B_0 , B_1 et B_2

g = 500									
g = 1 000									
g = 10 000									
Echantillon complet									
EEC									
EES									
Est. de B_0	53,3588	49,9532	52,6649	53,5876	56,7143	53,2401	56,3196		
Est. de B_1	0,3176	0,3251	0,3180	0,3171	0,3086	0,3179	0,3100		
Est. de B_2	-0,1326	-0,1258	-0,1302	-0,1330	-0,1378	-0,1324	-0,1377		
B_0 : Est. de la var.	416,1609	457,5178	293,8789	407,3107	224,0846	437,9610	251,3950		
B_1 : Est. var. $\times 10^{-3}$	2,1153	2,2925	1,1640	1,9127	0,5354	2,2366	0,8882		
B_2 : Est. var. $\times 10^{-3}$	2,7369	3,0352	2,4811	2,7226	2,3174	2,8028	2,3229		

6. CONCLUSION

Dans le présent article, nous avons présenté une théorie de l'échantillonnage inverse. L'efficacité de ce dernier augmente si l'on tire des sous-échantillons répétés, puis qu'on combine les résultats obtenus sur les sous-échantillons. Pour estimer un total, nous obtenons les conditions pour que l'estimateur limite sous échantillonnage inverse s'approche de l'estimateur sur l'échantillon complet (théorème 3) et pour que l'estimateur limite de la variance sous échantillonnage inverse s'approche de l'estimateur de la variance sous échantillon complet (théorème 4). Pour estimer des paramètres complexes, nous proposons une approche fondée sur des équations d'estimation combinées (BEC) et démontrons ses avantages par rapport à l'approche des équations d'estimation séparées (EES) (section 5). Nous étudions les algorithmes d'échantillonnage inverse pour certains plans d'échantillonnage à la section 2. Toutefois, des travaux supplémentaires seront nécessaires pour couvrir d'autres plans d'échantillonnage et pour éviter les limitations notées à la section 2. Nous sommes en train d'étudier diverses extensions en vue d'inclure les estimateurs sur échantillon complet stratifié à posteriori. L'analyse des données d'enquête nominales, les données en grappes sur la survie (Binder 1992) et les données d'enquête longitudinales.

REMERCIEMENTS

Les auteurs remercient le rédacteur adjoint et les examinateurs de leurs suggestions constructives. L'étude a été financée par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada.

Le résultat 1 découle directement de (3.1) en notant que, $s_0, \theta_1^*, \dots, \theta_g^*$ sont des variables aléatoires bornées indépendantes et identiquement distribuées (i.i.d.).

Le résultat 2 découle de la relation type entre les espérances conditionnelle et inconditionnelle :

$$E(\theta_g) = E[E(\theta_g | s_0)] = E[E(\theta_1^* | s_0)] = E(\theta_1^*).$$

Le résultat 3 découle du résultat correspondant pour les variances et de l'indépendance conditionnelle des θ_j^* sachant s_0 :

$$\text{Var}(\theta_g) = \text{Var}[E(\theta_g | s_0)] + E[\text{Var}(\theta_g | s_0)]$$
$$= \text{Var}(\theta_\infty) + \frac{1}{g} E[\text{Var}(\theta_1^* | s_0)].$$

Le résultat 4 découle directement du résultat 3.

Preuve du théorème 2

Le théorème 2 découle de l'application du résultat 3 du théorème 1 avec $g = 1$ pour obtenir

$$\text{Var}(\theta_\infty) = \text{Var}(\theta_1^*) - E[\text{Var}(\theta_1^* | s_0)],$$

puis de la substitution de cette expression à $\text{Var}(\theta_\infty)$ dans le résultat 3 du théorème 1 pour g en général.

Il découle maintenant de (5.21) - (5.23) que l'estimateur par linéarisation fondé sur les EBC (5.17) est identique à l'estimateur par échantillonnage inverse linéarisation de la variance (5.4).

Si nous passons à la régression linéaire avec $\mathbf{u}^k(\theta) = \mathbf{x}^k(\gamma^k - \mathbf{x}^k \theta)$, nous avons

$$\hat{\mathbf{V}}_{fu}^* = \frac{N_2}{N} \left(1 - \frac{m}{N} \right) \frac{1}{1}$$

$$\times \sum_{k \in s_j} \left[\mathbf{u}^k(\theta) - \bar{\mathbf{u}}_j(\theta) \right] \left[\mathbf{u}^k(\theta) - \bar{\mathbf{u}}_j(\theta) \right]^T, \quad (5.24)$$

où $\bar{\mathbf{u}}_j^k(\theta) = m_j^{-1} \sum_{k \in s_j} \mathbf{u}^k(\theta)$. En outre,

$$\hat{\mathbf{J}}_{gc}(\theta) = \frac{1}{N} \sum_{g=1}^G \frac{1}{m} \sum_{k \in s_g} \mathbf{x}^k \mathbf{x}^k_T$$

et

$$\hat{\mathbf{U}}_f^*(\theta) = \frac{m}{N} \sum_{k \in s_j} \mathbf{x}^k(\gamma^k - \mathbf{x}^k_T \theta).$$

Enfin, considérons le cas de la régression logistique avec $\mathbf{u}^k(\theta) = \mathbf{x}^k(\gamma^k - \mu^k(\theta))$. Dans ce cas, $\hat{\mathbf{V}}_{fu}^*$ est donné par (5.24) avec $\mathbf{u}^k(\theta) = \mathbf{x}^k(\gamma^k - \mu^k(\theta))$. En outre,

$$\hat{\mathbf{J}}_{gc}(\theta) = \frac{1}{N} \sum_{g=1}^G \frac{1}{m} \sum_{k \in s_g} \mu^k(\theta) (1 - \mu^k(\theta)) \mathbf{x}^k \mathbf{x}^k_T$$

et

$$\hat{\mathbf{U}}_f^*(\theta) = \frac{m}{N} \sum_{k \in s_j} \mathbf{x}^k(\gamma^k - \mu^k(\theta)).$$

De nouveau, il est important de noter que l'estimateur $\hat{\mathbf{V}}_{fu}^{gc}$ et l'estimateur de la covariance associée $\hat{\mathbf{V}}_L(\theta^{gc})$ peuvent être appliqués à partir d'un ensemble de microdonnées avec des données provenant de g sous-échantillons chacun de taille m . Ni les poids de sondage w_k ni les identificateurs de grappe ne sont nécessaires, si bien que le caractère confidentiel des microdonnées peut être préservé.

5.3 Exemple

Nous utilisons maintenant un ensemble de données présenté dans Batesse, Harter et Fuller (1988) pour illustrer les propriétés des méthodes fondées sur les équations d'estimation séparées et combinées. Les données ont été recueillies dans $k = 12$ comtés du centre-nord de l'Iowa. Les comtés ont été divisés en secteurs et un échantillon de secteurs a été sélectionné dans chaque comté. Ici, les comtés représentent les grappes et les secteurs échantillonnés dans un comté représentent les unités. Le nombre de secteurs échantillonnés (m_j) varie de 1 à 5 ce qui donne un total de $n = \sum_{j=1}^J m_j = 37$ unités échantillonnées. Pour

chaque unité échantillonnée (i, j), Batesse et coll. (1988) donnent le nombre d'hectares déclarés de maïs (y_{ij}^m) obtenus en interviewant les exploitants agricoles, ainsi que les nombres de pixels classés comme étant du maïs (x_{ij}^{1m}) et du soja (x_{ij}^{2m}) après des lectures faites par satellite de télédétection ($j = 1, \dots, m_j; i = 1, \dots, k$). Les données concernant l'un des secteurs échantillonnés paraissaient erronées et ont donc été exclues de l'analyse. Par conséquent, nous avons $n = 36$ observations (y_{ij}^m, x_{ij}^{1m}). À titre d'illustration, nous traitons l'échantillon comme s'il avait été sélectionné conformément au plan d'échantillonnage en grappes à deux degrés suivants : i) à la première phase, les comtés ont été sélectionnés avec remise et avec probabilité proportionnelle au nombre de secteurs M_j dans le comté; ii) à la deuxième phase, les secteurs échantillonnés ont été sélectionnés par échantillonnage aléatoire simple sans remise dans chaque comté sélectionné. Nous considérons deux paramètres, à savoir i) le ratio de population $\theta = R = Y/X$, où Y et X sont les totaux de la population de y et x , et ii) le coefficient de la régression « en cas de recensement » de y sur x , $\theta = B = (\sum_{i \in U} \mathbf{x}_i \mathbf{x}_i^T)^{-1} (\sum_{i \in U} \mathbf{x}_i y_i)$, où $\mathbf{x}_i = (1, x_{i1}^m, x_{i2}^m)^T$ et l représente une unité de population.

Pour certaines valeurs de g , nous générons g échantillons inverses indépendamment selon la procédure décrite pour le cas 2 à la section 2.3. Puis, nous utilisons les g sous-échantillons pour estimer R en utilisant la méthode des équations d'estimation séparées (EBS) et celle des équations d'estimation combinées (EBC) données à la section 5. Nous avons calculé les estimations correspondantes de la variance et les estimations par linéarisation de la variance des estimations sur l'échantillon complet θ .

Le tableau 1 donne l'estimation sur l'échantillon complet, R , l'estimation EBS R_g , l'estimation EBC R_{gc} et les estimations correspondantes de la variance. Il est évident, lorsqu'on examine le tableau 1, que les méthodes EBC et EBS donnent toutes deux de bons résultats comparativement à l'estimation sur l'échantillon complet R et à l'estimation correspondante de la variance par linéarisation sur l'échantillon complet, même pour $g = 500$. Le tableau 2 donne les résultats pour les coefficients de régression $\mathbf{B} = (B_0, B_1, B_2)^T$. À mesure que g augmente, les méthodes EBS et EBC semblent toutes deux concorder avec les estimations sur l'échantillon complet B_1 et B_2 , tandis que la méthode EBS donne une valeur un peu plus grande pour B_0 . Cependant, l'estimation de la variance par la méthode EBS donne de mauvais résultats, même pour la très grande valeur $g = 10\,000$, comparativement à l'estimation de la variance par linéarisation sur l'échantillon complet, la valeur par la méthode EBS étant égale à environ la moitié de la valeur correspondante sur l'échantillon complet pour B_0 et B_1 . Par contre, les estimations de la variance par la méthode EBC concordent bien avec les estimations de la variance sur l'échantillon complet, ce qui confirme la théorie.

  la convergence, nous obtenons l'estimateur EBC, $\hat{\mathbf{J}}_g(\hat{\theta}_{gc})$, ainsi que la matrice de l'information observ e, $\hat{\mathbf{J}}_g(\hat{\theta}_{gc})$. Notons que nous r solons les  quations d'estimation combin es (5.14) une seule fois pour obtenir $\hat{\theta}_{gc}$, contrairement   la m thode bas e sur les  quations d'estimation s par es qui n cessite la r solution des g  quations (5.10) pour obtenir $\hat{\theta}_1^*, \dots, \hat{\theta}_g^*$ et $\hat{\theta}_g^* = \sum_{j=1}^g \hat{\theta}_j^*/g$. Pour illustrer la m thode EBC propos e, consid rons le cas particulier du ratio $N^* = R$, pour lequel $u_k(\theta) = y_k - \theta x_k$. Les  quations d'estimation combin es (5.14) se r duisent   $Y^* - \theta X^* = 0$ et la solution $\hat{\theta}_{gc}$ est identique   l'estimateur sous  chantillonnage inverse combin  R_{gc} donn  par (5.1). En supposant qu'il y ait concordance avec l'EAS pour le premier moment, il d coule de (5.14) que $\hat{\theta}_{\infty}$ est une solution de

$$\hat{\mathbf{U}}_{\infty}(\theta) = E[\hat{\mathbf{U}}_1^*(\theta) | x_0] = \hat{\mathbf{U}}(\theta) = \mathbf{0}. \quad (5.16)$$

Par cons quent, $\hat{\theta}_{\infty} = \hat{\theta}$, ind pendamment de la taille m du sous- chantillon. Donc, le biais de $\hat{\theta}_{gc}$ est de m me ordre que le biais de $\hat{\theta}$ si g est suffisamment grand, ind pendamment de la taille m du sous- chantillon.

Nous appliquons maintenant la m thode de Binder (1983)   $\hat{\mathbf{U}}_{gc}(\hat{\theta}_{gc})$ pour obtenir un estimateur par lin arisation sous  chantillonnage inverse de $\mathbf{V}(\hat{\theta}_{gc})$. Il d coule de (5.8) que

$$\mathbf{V}^L(\hat{\theta}_{gc}) = [\mathbf{J}_g(\hat{\theta}_{gc})]^{-1} \mathbf{V}[\hat{\mathbf{U}}_{gc}(\hat{\theta}_{gc})] [\mathbf{J}_g(\hat{\theta}_{gc})]^{-1}, \quad (5.17)$$

  $\hat{\mathbf{V}}[\hat{\mathbf{U}}_{gc}(\hat{\theta}_{gc})]$ est l'estimateur de la variance du total estim , $\hat{\mathbf{U}}_{gc}(\hat{\theta})$, des $\mathbf{u}_k(\theta)$  valu    $\hat{\theta} = \hat{\theta}_{gc}$. Notons que $\hat{\mathbf{J}}_g(\hat{\theta}_{gc})$ est obtenue   la convergence de l'algorithme N-R appliqu    (5.14).

Puisque $\hat{\mathbf{U}}_{gc}(\hat{\theta})$ est l'estimateur sous  chantillonnage inverse du total $\mathbf{U}(\theta)$, il s'ensuit que l'estimateur sous  chantillonnage inverse de $\mathbf{V}[\hat{\mathbf{U}}_{gc}(\hat{\theta})]$ est donn  par

$$\hat{\mathbf{V}}_{gc}^{JU} = \frac{1}{g} \sum_{j=1}^g \hat{\mathbf{V}}_{jU}^*$$

$$- \frac{1}{g} \sum_{j=1}^g [\hat{\mathbf{U}}_j^*(\hat{\theta}) - \hat{\mathbf{U}}_{gc}(\hat{\theta})] [\hat{\mathbf{U}}_j^*(\hat{\theta}) - \hat{\mathbf{U}}_{gc}(\hat{\theta})]^T, \quad (5.18)$$

  $\hat{\mathbf{V}}_{gc}^{JU}$ est l'estimateur de la variance sous EAS pour le $j^{\text{ }}$ sous- chantillon, en supposant qu'il y ait concordance des deuxi mes moments. Si la concordance   trait   l'EAS sans remise, alors

$$\hat{\mathbf{V}}[\hat{\mathbf{U}}_{gc}(\hat{\theta}_{gc})] = \frac{1}{g} \sum_{j=1}^g \hat{\mathbf{V}}_{jU}^* - \frac{1}{g} \sum_{j=1}^g \hat{\mathbf{U}}_j^*(\hat{\theta}_{gc}) \hat{\mathbf{U}}_j^*(\hat{\theta}_{gc})^T = \hat{\mathbf{V}}^{JU}, \quad (5.20)$$

  $\hat{\mathbf{V}}^*$ est tir  de (5.19) par substitution de $\hat{\theta}_{gc}$   θ . Notons que $\hat{\mathbf{U}}_{gc}(\hat{\theta}_{gc}) = \mathbf{0}$.

Sous concordance des deuxi mes moments dans le cas de l'EAS, quand $g \rightarrow \infty$, il est facile de v rifier que $\hat{\mathbf{V}}^L(\hat{\theta}_{gc})$ converge vers l'estimateur de Binder $\hat{\mathbf{V}}^L(\hat{\theta})$ donn  par (5.8). Ceci s'obtient en notant que $\hat{\theta}_{\infty} = \hat{\theta}$, $\hat{\mathbf{J}}_g(\hat{\theta})$ et $\hat{\mathbf{V}}_{\infty}^{JU} = \mathbf{V}[\hat{\mathbf{U}}(\hat{\theta})]$ sous concordance des deuxi mes moments pour l'EAS. Donc, l'estimateur de la covariance $\hat{\mathbf{V}}^L(\hat{\theta}_{gc})$ donne des inf rences valides   sujet de θ pour un grand nombre de sous- chantillons, g , ind pendamment de la taille, m , du sous- chantillon.

Pour illustrer le calcul de l'estimateur par lin arisation sous  chantillonnage inverse, $\hat{\mathbf{V}}^L(\hat{\theta}_{gc})$, donn  par (5.17), consid rons le cas particulier d'un ratio $N^* = R$ avec $u_k(\theta) = y_k - \theta x_k$. Nous avons

$$\hat{\mathbf{V}}_{jU}^* = \frac{m}{N^2} \left(1 - \frac{m}{N} \right) \frac{1}{\sum_{k \in s_j^*} u_k(\theta) (\theta) - \bar{u}_j^*(\theta)} \left[\sum_{k \in s_j^*} u_k(\theta) (\theta) - \bar{u}_j^*(\theta) \right]^2, \quad (5.21)$$

  $\bar{u}_j^*(\theta) = \bar{y}_j^* - \theta \bar{x}_j^*$ et $(\bar{y}_j^*, \bar{x}_j^*)$ sont les moyennes du $j^{\text{ }}$ sous- chantillon. En outre,

$$f_{gc}^j(\theta) = \frac{N}{g} \sum_{j=1}^g \bar{x}_j^* = \bar{X}_g \quad (5.22)$$

et

$$\hat{\mathbf{U}}_j^*(\theta) = N(\bar{y}_j^* - \theta \bar{x}_j^*). \quad (5.23)$$

choisis de façon appropriée (Binder 1983; Godambe et Thompson 1986). Par exemple, considérons le cas scalaire de (5.5) en posant que $u_k^N(\theta) = y_k - \theta$. Ceci nous donne la moyenne de population $\theta_N^N = Y$. Parallelement, en posant que $u_k^N(\theta) = y_k - \theta x_k$ nous obtenons le ratio des totaux, $\theta_N^N = R = Y/X$. Le choix de $u^N(\theta) = x^N(y_k - \mu^N(\theta))$ avec $\mu^N(\theta) = x_T^N \theta$ donne les paramètres de régression linéaire « en cas de recensement »

$$\theta_N^N = \left(\sum_{k \in U} x_k x_k^T \right)^{-1} \sum_{k \in U} x_k y_k$$

Le choix de $u_k^T(\theta) = x^k(y_k - \mu^k(\theta))$ avec $\mu^k(\theta) = \log[\mu^k(\theta)/(1 - \mu^k(\theta))] = x^k \theta$ donne les paramètres de régression logistique « en cas de recensement » θ^N .

d'estimation, $u^k(\theta)$, qui produisent diverses mesures de l'incertitude du revenu, comme l'indice de Gini et l'indice de polarisation.

$$(5.6) \quad U(\theta) = \sum_{k \in s_0} w_k u^k(\theta) = 0,$$

où w_k est le poids de sondage lié à $k \in s_0$; en particulier, $w_k = 1/w_k$ si l'on utilise l'estimateur H-T de $U(\theta)$. La solution de (5.6) donne l'estimateur sur l'échantillon complet $\hat{\theta}$ qui, en général, est non linéaire et par conséquent, biaisé. Nous supposons que la taille de l'échantillon original, s_0 , est suffisamment grande pour pouvoir ignorer le biais de $\hat{\theta}$. Pour la régression logistique et d'autres cas complexes, il est nécessaire de résoudre (5.6) itérativement pour obtenir l'estimateur sur l'échantillon complet $\hat{\theta}$. Habituellement, on utilise pour cela l'algorithme de Newton-Raphson (N-R). La r^e étape de l'algorithme N-R est donnée par

$$(5.7) \quad \theta^{(r)} = \theta^{(r-1)} + J^{-1}(\theta^{(r-1)}) U(\theta^{(r-1)}),$$

où $\theta^{(r-1)}$ est la valeur de $\hat{\theta}$ obtenue à la $(r-1)^e$ itération et $U(\theta^{(r-1)})$ et $J(\theta^{(r-1)})$ sont les valeurs de $U(\theta)$ et $J(\theta) = -\partial^2 U(\theta)/\partial \theta^T = -\sum_{k \in s_0} w_k u^k(\theta)/\partial \theta^T$ évaluées à $\theta = \theta^{(r-1)}$. L'itération de l'algorithme N-R jusqu'à la convergence produit l'estimateur $\hat{\theta}$, ainsi que la matrice d'information observée $J(\hat{\theta})$.

Sous des conditions de régularité, Binder (1983) a obtenu un estimateur par linéarisation de Taylor de la matrice des covariances, $V(\hat{\theta})$, de $\hat{\theta}$ ayant la forme

$$(5.8) \quad V^L(\hat{\theta}) = [J(\hat{\theta})]^{-1} V[U(\hat{\theta})] [J(\hat{\theta})]^{-1},$$

où $V[U(\hat{\theta})]$ est un estimateur de la variance du total estimé, $U(\hat{\theta})$, des $u^k(\hat{\theta})$ évalué à $\theta = \hat{\theta}$. Par exemple, si $u^k(\hat{\theta}) = y_k - \theta x_k$, alors $\hat{\theta} = \sum_{k \in s_0} w_k y_k / \sum_{k \in s_0} w_k x_k = Y/X$ est l'estimateur du ratio et (5.8) se réduit à l'estimateur de la variance par linéarisation habituelle

$$(5.9) \quad V^L(\hat{\theta}) = \frac{1}{V^2} V \left[\sum_{k \in s_0} w_k u^k(\hat{\theta}) \right],$$

en notant que $J(\hat{\theta}) = \sum_{k \in s_0} w_k x_k x_k^T = X$.

(ii) Equations d'estimation séparées

Les estimateurs sous échantillonnage inverse individuels $\hat{\theta}_j^*$, $j = 1, \dots, g$ sont obtenus en résolvant les équations d'estimations séparées (EES)

$$(5.10) \quad U_j^*(\theta) = \sum_{k \in s_j} m \frac{1}{N} u^k(\theta) = 0; j = 1, \dots, g.$$

En général, nous avons besoin de g solutions itératives pour obtenir $\hat{\theta}_1^*, \dots, \hat{\theta}_g^*$. L'estimateur sous échantillonnage inverse de θ est alors donné par

$$(5.11) \quad \hat{\theta} = \frac{1}{g} \sum_{j=1}^g \hat{\theta}_j^*.$$

Il découle de (5.11) que $\hat{\theta}_\infty = E(\hat{\theta}_1^* | s_0)$ et $E(\hat{\theta}_\infty) = E(\hat{\theta}_1^*)$. En supposant qu'il y a concordance avec l'EAS pour le premier moment, il découle de (5.10) que le biais de $E(\hat{\theta}_1^*) - \theta$ est d'ordre m^{-1} , où m est la taille du sous-échantillon. L'estimateur sous échantillonnage inverse de $V(\hat{\theta})$ est donné par

$$(5.12) \quad V^g = \frac{1}{g} \sum_{j=1}^g V_j^* - \frac{1}{g} \sum_{j=1}^g V_j^* (\hat{\theta}_j^* - \theta) (\hat{\theta}_j^* - \theta)^T,$$

où V_j^* est donné par

$$(5.13) \quad V_j^* = [J_j^*(\hat{\theta}_j^*)]^{-1} V[U_j^*(\hat{\theta}_j^*)] [J_j^*(\hat{\theta}_j^*)]^{-1},$$

$V[U_j^*(\hat{\theta}_j^*)]$ est l'estimateur de la variance du total du j^e sous-échantillon $U_j^*(\hat{\theta}_j^*)$, représenté par V_j^{*T} (voir l'équation (5.19) plus loin), évalué à $\theta = \hat{\theta}_j^*$, et $J_j^*(\hat{\theta}_j^*)$ est

$$J_j^*(\hat{\theta}_j^*) = -\partial U_j^*(\hat{\theta}_j^*)/\partial \theta^T$$

Equations d'estimation combinées

Nous obtenons maintenant un estimateur par équations d'estimation combinées (EBC) $\hat{\theta}^{gc}$ qui donne des inférences valides quelle que soit la taille m du sous-échantillon. Nous combinons simplement les g équations dans (5.10) avant de les résoudre pour obtenir $\hat{\theta}$. Ceci nous donne les équations d'estimation combinées

$$(5.14) \quad U^{gc}(\theta) = \frac{1}{g} \sum_{j=1}^g U_j^*(\theta) = \frac{1}{g} \sum_{j=1}^g \sum_{k \in s_j} m \frac{1}{N} u^k(\theta) = 0.$$

En général, nous résolvons (5.14) en utilisant les itérations N-R (5.7) avec remplacement de $U(\theta^{(r-1)})$ par $U^{gc}(\theta^{(r-1)})$ et de $J(\theta^{(r-1)})$ par $J^{gc}(\theta^{(r-1)})$, où

$$(5.15) \quad J^{gc}(\theta) = -\frac{\partial U^{gc}(\theta)}{\partial \theta^T} = -\frac{1}{g} \sum_{j=1}^g \sum_{k \in s_j} m \frac{1}{N} \frac{\partial u^k(\theta)}{\partial \theta^T}.$$

5. APPROCHE DES ÉQUATIONS D'ESTIMATIONS COMBINÉES

À la présente section, nous étudions une approche de l'échantillonnage inverse fondée sur des équations d'estimation qui permet de faire des inférences valides au sujet de paramètres non linéaires, comme les ratios et les coefficients de régression linéaire ou logistique « en cas de section 3. L'estimateur sous échantillonnage inverse $\hat{\theta}_g^*$ donné par (3.1), à exactement le même biais que l'estimateur sur le sous-échantillon $\hat{\theta}_1^*$, et que le biais de $\hat{\theta}_1^*$ est d'ordre m^{-1} , où m est la taille du sous-échantillon. Par conséquent, le biais de $\hat{\theta}_g^*$ peut être appréciable, car m est habituellement beaucoup plus petit que la taille de l'échantillon original n . En fait, m peut être aussi petit que 2 pour les plans d'échantillonnage en grappes à deux degrés stratifiés avec deux grappes échantillonnées dans chaque strate. En outre, pour la régression logistique et d'autres cas, le calcul de $\hat{\theta}_g^*$ et de θ nécessite des solutions itératives. Par conséquent, l'application de $\hat{\theta}_g^*$ et de l'estimateur de la variance sous échantillonnage inverse \hat{V}_g^* , donné par (3.3), pourrait demander des calculs fastidieux quand le nombre d'échantillons inverses, g , est grand. Nous évitons ces difficultés en utilisant une approche fondée sur des équations d'estimations combinées (EEC).

À la section 5.1, nous considérons le cas spécial d'un ratio de totaux, $R = Y/X$, et expliquons l'« approche combinée » proposée par HOS vers la fin de la section 3.1. À la section 5.2, nous exposons la théorie générale et discussions de cas spéciaux. À la section 5.3, nous appliquons les résultats de la section 5.2 à un ensemble de données corréliées en grappes décrit dans Battese, Harter et Fuller (1988).

5.1 Ratio de totaux

$$R^{gc} = \frac{Y^g}{X^g}, \quad (5.1)$$

où $Y^g = g^{-1} \sum_{j=1}^g Y_j^*$ et $X^g = g^{-1} \sum_{j=1}^g X_j^*$. Maintenant, en supposant que la taille finale de l'échantillon « combiné » soit suffisamment grande, il découle de (5.1) que

$$E(R^{gc}) \approx \frac{E(Y^g)}{E(X^g)} = \frac{Y}{X} = R$$

sous les conditions du théorème 3. Autrement dit, R^{gc} est approximativement sans biais pour R , indépendamment de la taille du sous-échantillon, à condition que g soit suffisamment grand. Par conséquent, en utilisant l'approximation par linéarisation de Taylor, nous obtenons pour la variance de R^{gc} l'expression

$$V(R^{gc}) \approx \frac{1}{X^2} V(U^g), \quad (5.2)$$

où $\hat{U}^g = g^{-1} \sum_{j=1}^g \hat{U}_j^*$ est l'estimateur sous échantillonnage inverse du total U des résidus $u_i = y_i - Rx_i$, $i = 1, \dots, N$. En notant que \hat{U}_j^* est l'estimateur sous échantillonnage inverse d'un total, il découle de (3.3) qu'un estimateur sous échantillonnage inverse de $V(U^g)$ est donné par

$$\hat{V}^{gu} = \frac{1}{g} \sum_{j=1}^g \hat{V}_j^{gu*} - \frac{1}{g} \sum_{j=1}^g \hat{U}_j^* (\hat{U}_j^* - \hat{U}^g)^2, \quad (5.3)$$

où \hat{V}_j^{gu*} est l'estimateur de la variance produit d'après le j^{e} sous-échantillon. Puisque R est inconnu, nous le remplaçons par R^{gc} dans (5.3) pour obtenir l'estimateur de la variance \hat{V}^{gu} . Maintenant, en remplaçant X par son estimateur X^{gu} et $V(U^g)$ par \hat{V}^{gu} dans (5.2), nous obtenons pour l'estimateur de la variance par linéarisation sous échantillonnage inverse de R^{gc} l'expression

$$\hat{V}^L(R^{gc}) = \frac{1}{X^2} \hat{V}^{gu}. \quad (5.4)$$

Sous les conditions du théorème 4, $\hat{V}^L(R^{gc})$ converge vers l'estimateur de la variance par linéarisation habituel de l'estimateur sur l'échantillon complet $\hat{R} = Y/X$.

5.2 Paramètres non linéaires

1) Équations d'estimation sur l'échantillon complet

On pourrait considérer un vecteur de paramètres en population finie θ_N comme étant la solution des équations d'estimation (EB) « en cas de recensement » :

$$U(\theta) = \sum_{k \in U} u^k(\theta) = 0, \quad (5.5)$$

où $\sum_{k \in U}$ représente la sommation sur la population finie U de taille N , et où les fonctions d'estimation $u^k(\theta)$ sont

L'estimateur de Y sur le j^{e} échantillon inverse. Il est facile de vérifier que $\hat{Y} = \hat{Y}^{\text{pps}}$, et que $\hat{V} = \hat{V}^{\text{pps}}$, en notant que $Y_i^* = (N/k) \sum_{i \in s_j} Y_i^i$ où Y_i^i représente la valeur de l'unité d'un échantillon inverse sélectionnée à partir de la grappe lors du i^{e} tirage. Donc, l'échantillonnage inverse préserve à la fois l'estimateur et l'estimateur de la variance.

ii) Échantillonnage en grappes à deux degrés

Partant du cas de grappes de taille inégale, M_i , nous sélectionnons les grappes avec PPT et avec remise, puis nous tirons des sous-échantillons aléatoires simples de taille

égale, m , indépendamment dans chaque grappe échantillonnée avec remise. L'estimateur de Y est $\hat{Y}^{\text{pps}} = (N/k) \sum_{i=1}^I \hat{Y}_i^i$ où \hat{Y}_i^i est la moyenne d'échantillon de la grappe sélectionnée lors du i^{e} tirage. L'estimateur de la variance de \hat{Y}^{pps} est donné par

$$\hat{V}^{\text{pps}} = N^2 \frac{1}{k} \sum_{i=1}^{I-1} \left(\hat{Y}_i^i - \frac{1}{k} \sum_{i=1}^I \hat{Y}_i^i \right)^2.$$

L'estimateur sous échantillonnage inverse est donné par $\hat{Y}^g = \delta^{-1} \sum_{i=1}^I \hat{Y}_i^g$, où $Y_i^* = (N/k) \sum_{i \in s_j} Y_i^i$, et Y_i^i est défini comme plus haut. Il est facile de vérifier que $\hat{Y} = \hat{Y}^{\text{pps}}$ et que $\hat{V} = \hat{V}^{\text{pps}}$. Donc, l'échantillonnage inverse préserve à la fois l'estimateur et l'estimateur de la variance.

4.3 Concordance approximative

À la section 2, nous avons fait remarquer que la concordance exacte avec l'EAS est difficile à obtenir lorsque le plan d'échantillonnage original comprend des grappes. Nous proposons plusieurs méthodes donnant une concordance approximative pour surmonter cette difficulté. À la présente sous-section, nous étudions les propriétés des méthodes de concordance approximative.

4.3.1 Échantillonnage en grappes à un degré

À la section 2.2, Cas 1, nous considérons le cas de grappes de taille égale, M , et de l'échantillonnage aléatoire simple de grappes. L'estimateur d'un total Y est donné par $\hat{Y} = (K/k) \sum_{i=1}^k Y_i$, où Y_i est le total dans la i^{e} grappe échantillonnée et K est le nombre de grappes dans la population. L'estimateur de la variance de \hat{Y} est

$$\hat{V} = K^2 \left(\frac{1}{k} - \frac{1}{K} \right) \sum_{i=1}^{k-1} \left[Y_i - \frac{1}{k} \sum_{i=1}^k Y_i \right]^2.$$

Pour l'échantillonnage inverse, nous avons proposé d'obtenir une concordance approximative en sélectionnant une unité au hasard dans chaque grappe échantillonnée, $i (= 1, \dots, k)$. L'estimateur sous échantillonnage inverse est donné par $\hat{Y}^g = \delta^{-1} \sum_{i=1}^I Y_i^*$ où $Y_i^* = N Y_i^i$ représente l'estimateur du total Y pour le j^{e} échantillon inverse. L'estimateur de la variance sous échantillonnage inverse, \hat{V}^g , est donné par (3.4).

$$\frac{\hat{V}^g}{\hat{V}} = 1 - \frac{K}{k} \left[1 - \left(1 - \frac{K}{m} \right) \frac{1}{k} \frac{M}{s_{1y}^2} \frac{s_{2y}^2}{s_{1y}^2} \right]$$

quand $g \rightarrow \infty$. Il découle de (4.8) et (4.9) que

$$(4.9) \quad \hat{V}^g = N^2 \frac{1}{s^2} k \approx 1 - \frac{K}{k},$$

Il est facile de vérifier que $\hat{Y}^g = Y$ de sorte que la concordance approximative préserve l'estimateur original Y à la limite. Par ailleurs, nous pouvons montrer que \hat{V}^g tend vers

(3.4). La variance sous échantillonnage inverse, \hat{V}^g , est donné par $\hat{Y}^g = \delta^{-1} \sum_{i=1}^I Y_i^*$, où $Y_i^* = (N/k) \sum_{i \in s_j} Y_i^i$. L'estimateur sous échantillonnage inverse du total Y est donné par \hat{Y}^g . Représentons les valeurs des unités par Y_1^i, \dots, Y_k^i . L'estimateur dans chaque grappe échantillonnée $i (= 1, \dots, k)$. Représentons les valeurs des unités échantillonnées d'une unité au hasard à partir des m unités échantillonnées d'obtenir une concordance approximative en sélectionnant Pour l'échantillonnage inverse, nous avons proposé Cochran 1977, pages 276-278).

$$(4.8) \quad \hat{V} = N^2 \left\{ \frac{1}{k} \left(1 - \frac{K}{k} \right) s_{1y}^2 + \frac{K}{k} \left(1 - \frac{m}{m} \right) \frac{1}{s_{2y}^2} \right\},$$

est donné par L'estimateur de la variance de Y à la grappe échantillonnée. L'estimateur de l'échantillon de $\sum_{i=1}^I Y_i^*$, où $Y_i^* = M Y_i^i$ et Y_i^i est la moyenne d'échantillon de l'estimateur H-T du total Y est donné par $\hat{Y} = (K/k)$ taille égale, M , et EAS sans remise aux deux étapes. L'échantillonnage en grappes à deux degrés avec grappes de

4.3.2 Échantillonnage en grappes à deux degrés

À la section 2.3, Cas 1, nous considérons le cas de l'échantillonnage en grappes à deux degrés avec grappes de taille égale, M , et EAS sans remise aux deux étapes. L'estimateur H-T du total Y est donné par $\hat{Y} = (K/k)$ taille égale, M , et EAS sans remise aux deux étapes. L'estimateur de la variance de Y est donné par

$$(4.7) \quad \hat{V}/\hat{V}^g = 1 - K/k.$$

Il est facile de vérifier que $\hat{Y}^g = Y$, si bien que la concordance approximative préserve l'estimateur original Y à la limite. Par ailleurs, nous pouvons montrer que

Il s'ensuit que \hat{V}^g donne de nouveau une surestimation de la variance si la fraction d'échantillonnage k/K n'est pas faible.

Théorème 3

Soit $\pi_i(s_0)$ la probabilité conditionnelle que la i^{e} unité soit sélectionnée dans le sous-échantillon pour un échantillon initial donné, s_0^* . Supposons que $\theta_j^* = Y_j^*$ est l'estimateur H-T d'un total $\theta = Y$ pour le j^{e} sous-échantillon. Alors, l'estimateur limite sous échantillonnage inverse, $\theta_\infty^* = Y_\infty^*$, sera l'estimateur H-T, Y_j , sous le plan de sondage original si, et uniquement si, les probabilités conditionnelles d'inclusion $\pi_i(s_0)$ sont constantes pour tous les s_0 contenant la i^{e} unité, c'est-à-dire $\pi_i(s_0) = \pi_i$ pour tout $s_0^* \supset i$.

La condition $\pi_i(s_0) = \pi_i$ est assez naturelle pour la plupart des plans de sondage pour lesquels l'estimateur H-T est utilisé. Si les sous-échantillons sont tous des échantillons aléatoires simples de taille fixe m , alors l'estimateur pour un sous-échantillon est simplement N_j^*/N_j^* qui est l'estimateur naturel sous échantillonnage aléatoire simple.

Le théorème 4 qui suit établit les conditions sous lesquelles l'estimateur de la variance sous échantillonnage inverse, $V_{g,HT}^*$, de Y_g^* converge vers V_{HT}^* , c'est-à-dire l'estimateur H-T de la variance sur l'échantillon complet Y , quand $g \rightarrow \infty$. Nous avons

$$V_{HT}^* = \sum_{i \in s_0} \frac{\pi_{ii}^* \pi_l^* \pi_l^*}{\pi_{ii}^* - \pi_l^*} Y_i Y_l \quad (4.2)$$

(voir Cochran 1977, page 261) et

$$V_{g,HT}^* = \frac{1}{g} \sum_{j=1}^g V_{HT}^* - \frac{1}{g} \sum_{j=1}^g (Y_j^* - \bar{Y}_g^*)^2 \quad (4.3)$$

avec

où π_{ii}^* est la probabilité d'inclusion conjointe inconditionnelle des i et l unités. Si le sous-échantillon s_j^* est inconditionnellement un sous-échantillon aléatoire simple, alors $\pi_{ii}^* = m(m-1)/(N(N-1))$, $i \neq l$. Notons que $V_{j,HT}^*$ est l'estimateur H-T de la variance de Y_j^* et que $\pi_{ii}^* = \pi_i^*$, $\pi_{ii}^* = \pi_i$.

Théorème 4

Si $V_{j,HT}^*$ est l'estimateur d'Horvitz-Thompson (H-T) de la variance de Y_j^* pour le j^{e} sous-échantillon, alors conditionnellement à s_0^* , $V_{g,HT}^*$ converge vers l'estimateur (H-T) de la variance de Y pour le plan de sondage original, quand $g \rightarrow \infty$, si les probabilités d'inclusions conjointes conditionnelles sont constantes pour tous les s_0 contenant une paire donnée (i, l) d'unités, c'est-à-dire $\pi_{ii}^*(s_0) = \pi_{ii}^*$ pour tout $s_0 \supset \{i, l\}$.

Dans le théorème 4, nous considérons l'estimateur H-T de la variance. Cependant, on préfère souvent l'estimateur d'Horvitz-Thompson (S-Y-G), V_{SYG}^* , à l'estimateur sous échantillonnage inverse correspondant à Y^{dps} est donné par $Y_g^* = g^{-1} \sum_{j=1}^g Y_j^*$, où Y_j^* représente

$$V^{\text{dps}} = N^2 \frac{k}{k-1} \sum_{i=1}^k \left(\bar{Y}_i' - \frac{1}{k} \sum_{l=1}^k \bar{Y}_l' \right)^2.$$

Y^{dps} est donné par

Dans le cas de l'échantillonnage PPT et avec remise de grappes de taille inégale, M_i , nous avons une concordance exacte avec l'EAS avec remise. Les estimations de Y sont données par $Y^{\text{dps}} = (N/k) \sum_{i=1}^k X_i'$, où N est le nombre total d'unités de la population et X_i' est la moyenne de la grappe sélectionnée lors du i^{e} tirage. L'estimateur Y^{dps} n'est pas égal à l'estimateur H-T de Y . L'estimateur de la variance de Y^{dps} est donné par

4.2 Concordance exacte : estimations PPT

L'estimateur de la variance sous échantillonnage inverse S-Y-G (4.4) quand $g \rightarrow \infty$.

Théorème 5

Le théorème 5 qui suit montre que $V_{g,SYG}^*$ ne converge pas vers V_{SYG}^* quand $g \rightarrow \infty$, c'est-à-dire $V_{g,SYG}^* \neq V_{SYG}^*$. Si le sous-échantillon est inconditionnellement un échantillon aléatoire simple, c'est-à-dire $\pi_i^* = m/N$ et $\pi_{ii}^* = m(m-1)/(N(N-1))$, $i \neq l$, alors $V_{j,HT}^* = V_{j,SYG}^*$ et la variance de Y_j , c'est-à-dire l'estimateur H-T de la

$$V_{g,SYG}^* = \frac{1}{g} \sum_{j=1}^g V_{SYG}^* - \frac{1}{g} \sum_{j=1}^g (Y_j^* - \bar{Y}_g^*)^2 \quad (4.6)$$

est donné par

$$V_{j,SYG}^* = \sum_{i < l \in s_j^*} \frac{\pi_{ii}^* \pi_l^* - \pi_{ii}^*}{\pi_l^* - \pi_i^*} Y_i Y_l \quad (4.5)$$

L'estimateur S-Y-G de la variance de Y_j^* est

$$V_{SYG}^* = \sum_{i < l \in s_0} \frac{\pi_{ii}^* \pi_l^* - \pi_{ii}^*}{\pi_l^* - \pi_i^*} Y_i Y_l, \quad (4.4)$$

donné par

L'estimateur H-T de la variance, V_{HT}^* , parce qu'il est plus stable et qu'on connaît plusieurs plans de sondage pour lesquels il est systématiquement non négatif, tandis que V_{HT}^* prend souvent des valeurs négatives (Cochran 1977, page 261). L'estimateur S-Y-G de la variance de Y existe pour les plans de sondage à taille d'échantillon fixe et est

(HOS) ont obtenu les mêmes résultats par simulation, mais cela n'est pas nécessaire compte tenu du résultat 4 du théorème 1). Nous voyons que $g = 100$ sous-échantillons conviendrait pour de nombreuses applications et que nous obtenons une efficacité presque totale avec $g = 1\ 000$.

Le fait que les $\theta_1^g, \dots, \theta_g^g$ ne soient pas indépendamment indépendants signifie que l'estimation de $\text{Var}(\theta^g)$ n'est pas tout à fait simple. Toutefois, on peut obtenir un estimateur assez simple de la variance en utilisant le théorème 2 qui suit.

Théorème 2

La variance de θ^g peut être exprimée sous la forme

$$\text{Var}(\theta^g) = \text{Var}(\theta_1^g) - \frac{g}{g-1} E[\text{Var}(\theta_1^g | s_0)] \quad (3.2)$$

Nous pouvons estimer le premier terme de (3.2) par \hat{V}_1^g pour $j = 1, \dots, g$, et donc par leur moyenne $g^{-1} \sum_{j=1}^g \hat{V}_j^g$. En outre, l'expression

$$s_{\theta^g}^2 = \frac{1}{g} \sum_{j=1}^g (\theta_j^g - \bar{\theta}^g)^2$$

donne un estimateur sans biais de $E[\text{Var}(\theta_1^g | s_0)]$, parce que les $\theta_1^g, \dots, \theta_g^g$ sont conditionnellement indépendants sachant l'échantillon initial, s_0 . Cela nous mène à un estimateur de $\text{Var}(\theta^g)$ de la forme

$$\hat{V}^g = \frac{1}{g} \sum_{j=1}^g \hat{V}_j^g - \frac{1}{g} \sum_{j=1}^g (\hat{\theta}_j^g - \bar{\theta}^g)^2 \quad (3.3)$$

Les propriétés de l'estimateur de la variance \hat{V}^g dépendent des propriétés de l'estimateur sur le sous-échantillon \hat{V}_j^g . Par exemple, si \hat{V}_j^g est sans biais, alors \hat{V}^g est également sans biais.

Pour le cas spécial d'un total de population $\theta = X$ et d'un sous-échantillonnage aléatoire simple, c'est-à-dire $p(s^*) = 1/\binom{N}{n}$, nous avons $\theta_j^g = X_j^g = N Y_j^g$ et X_j^g est sans biais pour X avec un estimateur de la variance sans biais $\hat{V}_j^g = N^2 (m^{-1} - N^{-1}) s_{Y_j^g}^2$, où \bar{Y}_j^g est la moyenne et $s_{Y_j^g}^2$ est la variance du j^{e} sous-échantillon. L'estimateur de la variance \hat{V}^g de $\theta^g = X^g = g^{-1} \sum_{j=1}^g (N \bar{Y}_j^g)$, donné par (3.3), est sans biais et il se réduit à

$$\hat{V}^g = \frac{1}{g} \sum_{j=1}^g \hat{V}_j^g - \frac{1}{N^2} \sum_{j=1}^g (\bar{Y}_j^g - \bar{Y}^g)^2 \quad (3.4)$$

où $\bar{Y}^g = g^{-1} \sum_{j=1}^g \bar{Y}_j^g$ est l'estimateur de la variance en commençant par exprimer $\text{Var}(X^g)$ sous la forme

$$\text{Var}(X^g) = N^2 \frac{m}{m-1} S_2^g + \frac{1}{g} \sum_{j=1}^g \text{Var}(Y_j^g) - N^2 \frac{mg}{m-1} E(s_{Y_j^g}^2) \quad (3.5)$$

où S_2^g est la variance de population et $s_{Y_j^g}^2$ est la variance d'échantillon en utilisant le nombre total gm d'unités

4.1 Équivalence exacte

4. ESTIMATION D'UN TOTAL

Il découle de (3.6) que l'estimateur de la variance de HOS est en fait identique à notre estimateur de la variance (3.4) et est également exactement sans biais.

$$\hat{V}^g = \quad (3.6)$$

$$\hat{V}^{g(\text{HOS})} = N^2 \left(\frac{m}{m-1} - \frac{1}{g} \sum_{j=1}^g s_{Y_j^g}^2 + \frac{1}{g} \sum_{j=1}^g \hat{V}_j^g \right) - N^2 \left(\frac{m}{m-1} - \frac{1}{g} \sum_{j=1}^g s_{Y_j^g}^2 + \frac{1}{g} \sum_{j=1}^g (\bar{Y}_j^g - \bar{Y}^g)^2 \right)$$

sous-échantillonnées. À la deuxième étape, ils notent qu'un peut générer un estimateur approximativement sans biais de $\text{Var}(X_j^g)$ à partir de (3.5) en remplaçant S_2^g et $\text{Var}(Y_j^g)$ par $s_{Y_j^g}^2$ et $E(s_{Y_j^g}^2)$ en remplaçant $E(s_{Y_j^g}^2)$ par $s_{Y_j^g}^2$. Nous suivons maintenant ce mode opératoire et obtenons une forme explicite de l'estimateur de la variance de HOS représentée par $\hat{V}^{g(\text{HOS})}$. En notant que chaque $s_{Y_j^g}^2$ est sans biais pour S_2^g , nous obtenons un estimateur agrégé sans biais de S_2^g de la forme $g^{-1} \sum_{j=1}^g s_{Y_j^g}^2$. En outre, $s_{Y_j^g}^2$ peut être décomposé selon $(mg - 1) s_{Y_j^g}^2 = (m - 1) \sum_{f=1}^{g-1} s_{Y_j^g}^2 + m \sum_{f=1}^{g-1} (\bar{Y}_j^g - \bar{Y}^g)^2$. Donc,

$$\hat{V}_j^g = \frac{1}{g} \sum_{f=1}^g \hat{V}_j^g$$

original, quand $g \rightarrow \infty$, où

$$(4.1)$$

Comme nous l'avons montré à la section 3, le sous-échantillonnage répété augmente l'efficacité d'un estimateur, mais cela ne signifie pas nécessairement que l'estimateur sous échantillonnage inverse, $\hat{\theta}^g$, converge vers l'estimateur sur l'échantillon complet original, θ , vers l'estimateur sur l'échantillon complet original, θ , quand $g \rightarrow \infty$, même si nous partons d'un estimateur sans biais pour le cas particulier d'un total $\theta = X$ et considérons l'estimateur sans biais d'Horvitz-Thompson (H-T), $\hat{Y} = \sum_{i \in s_g} Y_i / \pi_i$, fondé sur l'échantillon original complet. Le théorème 3 qui suit établit les conditions sous lesquelles l'estimateur sous échantillonnage inverse correspondait

et π_i^* est la probabilité d'inclusion inconditionnelle de la i^{e} unité. Si le sous-échantillon s_j^* est inconditionnellement un échantillon aléatoire simple, alors $\pi_i^* = m/N$, où m est la taille du sous-échantillon.

estimateur sous échantillonnage inverse « séparé » de θ fondé sur les g sous-échantillons est donné par

$$\hat{\theta}_g = \frac{1}{g} \sum_{j=1}^g \hat{\theta}_j. \quad (3.1)$$

Nous représentons l'estimateur fondé sur s_0 par $\hat{\theta}$. Le théorème 1 qui suit donne les résultats fondamentaux concernant $\hat{\theta}_g$ et sa variance.

Théorème 1
1. Conditionnellement à l'échantillon original, s_0 , $\hat{\theta}_g$ converge presque certainement vers $E(\hat{\theta}_1^* | s_0) = \theta_g^*$, disons, quand $g \rightarrow \infty$.

$$2. E(\hat{\theta}_g) = E(\hat{\theta}_1^*).$$

$$3. \text{Var}(\hat{\theta}_g) = \text{Var}(\hat{\theta}_1^*) + \frac{g}{1} E[\text{Var}(\hat{\theta}_1^* | s_0)].$$

$$4. \text{Si } r_g = \frac{\text{Var}(\hat{\theta}_g)}{\text{Var}(\hat{\theta}_1^*)}, \text{ alors } r_g = 1 + \frac{r_1}{r_1 - 1}.$$

Le résultat 4 du théorème 1 démontre que l'augmentation du nombre de sous-échantillons, g , augmente effectivement l'efficacité de $\hat{\theta}_g$. Plus précisément, le rapport des variances r_g prend la forme $a + b/g$. Si l'estimateur sur le sous-échantillon, $\hat{\theta}_1$, est sans biais pour θ , il en est de même pour l'estimateur en échantillonnage inverse, $\hat{\theta}_g^*$. Cependant, si $\hat{\theta}_1^*$ présente un biais d'ordre m^{-1} , où m représente la taille du sous-échantillon, alors $\hat{\theta}_g^*$ présente exactement le même biais. Puisque m est habituellement nettement plus petit que la taille de l'échantillon original, ce biais peut être appréciable. Il s'agit là d'une limitation sérieuse de $\hat{\theta}_g$ dans les cas non linéaires, tels que les ratios et les coefficients de régression. À la section 5, nous proposons un autre estimateur de θ fondé sur l'approche des équations d'estimation (EB). Cette estimateur est asymptotiquement sans biais pour toute valeur de m à mesure que la taille de s_0 augmente, contrairement à $\hat{\theta}_g^*$.

Le résultat 4 du théorème 1 peut être utilisé pour déterminer le nombre de sous-échantillons, g , nécessaires pour obtenir une efficacité raisonnable. Par exemple, HOS donne un exemple où $r_1 = 29.3$. L'échantillon original était un échantillon aléatoire stratifié très efficace contenant $n = 15\ 618$ observations tirées de la Statistics of Income corporate survey, tandis que le sous-échantillon était un échantillon aléatoire simple de $m = 2\ 224$ observations. Un sous-échantillon unique est assez inefficace. Cependant, dans ce cas, l'échantillonnage inverse répété permet de récupérer toute l'information contenue dans l'échantillon original à la limite. L'application du résultat 4 du théorème 1 aboutit directement au tableau qui suit :

g	1	10	3.83	1.28	1.03
r_g	29.3	10	3.83	1.28	1.03

Supposons que nous ayons un échantillon à deux degrés tirés dans chaque strate, où les grappes sont échantillonnées avec PPT et avec remise, et que le sous-échantillonnage est effectué indépendamment dans chaque grappe échantillonnée. En utilisant la méthode d'échantillonnage inverse du cas 2, section 2.3, nous obtenons des échantillons aléatoires simples provenant de chaque strate. Nous pouvons alors appliquer la méthode de la section 2.1, en traitant les échantillons inverses comme s'ils avaient été tirés sans remise pour obtenir un échantillon inverse de taille $k_0 = \min_h (k_h)$, où k_h est le nombre de grappes échantillonnées dans la strate h . Dans le cas important où $k_h = 2$ UPE sont échantillonnées dans chaque strate, la taille de l'échantillon inverse, k_0 , est seulement égale à deux.

2.4 Échantillonnage stratifié à deux degrés

Le scénario de sous-échantillonnage se déduit facilement. Il découle de (2.2) que $P(s^* | s_0) = \prod_{i=1}^k (1/m_i')$ et, donc,

$$\sum_{s_0 \in s^*} P_0(s_0) = \prod_{i=1}^k \left(\frac{M_i'}{M_i' - 1} \right) \left(\frac{N}{M_i' - 1} \right) \left(\frac{m_i'}{M_i' - 1} \right) = \prod_{i=1}^k \frac{m_i'}{M_i' - 1}.$$

Les résultats de la présente section sont assez généraux et s'appliquent aussi bien aux enquêtes par sondage qu'aux situations avec mise en grappes considérées par Hoffman et coll. (2001). Supposons que nous voulions estimer un paramètre de population, θ , et que nous ayons un échantillon, s_0 , d'observations tiré de la population selon un plan de sondage complexe. Supposons aussi que nous ayons un algorithme de sous-échantillonnage permettant de produire des échantillons à partir d'un plan de sondage plus simple. Souvent, il s'agit de l'échantillonnage aléatoire simple, mais nous pouvons étendre considérablement la gamme d'applications en permettant éventuellement des plans de (sous-) échantillonnage plus généraux, comme l'EAS stratifié si l'échantillon original est un échantillon à deux degrés stratifié. La seule exigence à laquelle doit satisfaire le plan de sondage plus simple est que nous puissions produire un estimateur du paramètre d'intérêt, θ , ainsi qu'un estimateur de sa variance. Représentons par $\hat{\theta}_j$ l'estimateur et par V_j l'estimateur de sa variance produits à partir du j^{e} sous-échantillon quand nous générons une série de g sous-échantillons conditionnellement indépendants s_j^* ($j = 1, \dots, g$). Notons que les $\hat{\theta}_j$ ne sont pas moyens sur la distribution de l'échantillon initial, s_0 . Un

grappes; 3) grappes de taille inégale, M_i , et échantillonnage des grappes avec probabilité proportionnelle à la taille M_i et avec remise.

Cas 1. Dans le cas de grappes de taille égale, M_i , et d'un échantillonnage aléatoire simple des grappes, il est difficile d'obtenir une concordance exacte avec l'EAS. Supposons que s_0 contienne k grappes tirées à partir de K grappes dans la population ($N = KM$). Une méthode approximative simple de sous-échantillonnage consiste à sélectionner une unité au hasard dans chaque grappe échantillonnée de sorte que la taille de s^* soit k . Hoffman, Sen et Weinberg (2001) utilisent une méthode comparable pour des applications biostatistiques. HOS utilisent l'échantillonnage systématique pour sélectionner un cas dans chaque grappe échantillonnée.

Cas 2. Hoffman et coll. (2001) sélectionnent une unité au hasard dans chaque grappe dans le cas de grappes de taille inégale, dans un cadre de travail fondé sur un modèle pour les données mises en grappes. Dans le cas d'applications d'échantillonnage, cette méthode ne marche pas en ce sens qu'il est impossible d'obtenir des échantillons aléatoires simples de taille fixe par sous-échantillonnage, même approximativement. HOS ont proposé une autre solution qui consiste à agrandir artificiellement la population pour retomber sur le cas de grappes de taille égale, puis à appliquer la méthode de sous-échantillonnage utilisée au cas 1. Pour commencer, nous forçons toutes les grappes à avoir la même taille en leur ajoutant un nombre approprié de pseudo-unités de sorte que leur taille soit égale à celle de la grappe échantillonnée la plus grande. Puis, nous tirons une unité au hasard dans chaque grappe échantillonnée et nous écartons toute pseudo-unité pour obtenir l'échantillon final. Cette méthode approximative fait dépendre $p(s|s_0)$ de s_0 parce que la probabilité conditionnelle dépend de la taille $M(s_0)$ de la grappe échantillonnée la plus grande.

Cas 3. Pour le cas des grappes de taille inégale M_i et de l'échantillonnage avec probabilité proportionnelle à la taille (PPT) et avec remise, HOS ont proposé une méthode simple de sous-échantillonnage qui donne $p(s^*) = (1/N)^k$, où s^* représente maintenant un échantillon aléatoire simple ordonné tiré avec remise à partir des $N = \sum_{i=1}^K M_i$ unités dans la population, c'est-à-dire $s^* = (i_1^*, \dots, i_k^*)$, où i_j^* représente l'unité obtenue lors du j^{e} tirage ($j = 1, \dots, k$). Considérant les grappes échantillonnées comme étant ordonnées, nous sélectionnons une unité au hasard dans chacune d'elles. Notons qu'une même grappe pourrait figurer plus d'une fois dans l'échantillon ordonné. Si nous représentons la taille de la grappe obtenue lors du i^{e} tirage PPT par $M_{i'}^t$, alors

$$p(s^*) = \left[\prod_{i=1}^k \frac{1}{M_{i'}^t} \right] \left[\prod_{k=1}^K \frac{1}{M_i^t} \right] = \left(\frac{1}{N} \right)^k, \quad (2.7)$$

où $\prod_{k=1}^K (M_i^t/N)$ est la probabilité de tirer l'échantillon ordonné de grappes. Notons que s_0 est l'échantillon PPT ordonné et que la sommation dans (2.1) ne compte qu'un seul terme.

Si les grappes sont tirées avec probabilité d'inclusion $\pi_i = kM_i/N$ et sans remise, alors il est impossible d'égaliser l'EAS. Cependant, nous pouvons traiter les grappes comme si elles étaient tirées avec remise, comme cela se fait en pratique, puis appliquer le scénario du cas 3. Il y aura surestimation de la variance si la variance de l'estimateur obtenu est plus faible que celle de l'estimateur sous-échantillonné PPT avec remise (voir, par exemple, Wolter 1985, page 45). Toutefois, la surestimation n'est pas grave si la traction d'échantillonnage k/K est faible (voir la section 4.3).

2.3 Échantillonnage en grappes à deux degrés

HOS ont également étudié l'échantillonnage à deux degrés pour les cas suivants : 1) grappes de taille égale, M_i , et k grappes échantillonnées avec probabilité égale à la première phase; sous-échantillons aléatoires simples de taille égale, m , tirés indépendamment dans chaque grappe échantillonnée (UPB) et 2) grappes de taille inégale, M_i , et k grappes échantillonnées avec PPT et avec remise; sous-échantillons aléatoires simples de taille inégale, m_i , tirés indépendamment dans chaque grappe échantillonnée avec remise.

Cas 1. Comme dans le cas de l'échantillonnage en grappes à un degré, il est difficile d'appliquer une méthode exacte d'échantillonnage inverse. Une méthode approximative simple d'échantillonnage inverse consiste à sélectionner une unité au hasard dans chacun des k sous-échantillons.

Cas 2. Comme dans le cas 3 d'échantillonnage en grappes à un degré, nous sélectionnons simplement une unité au hasard dans chacun des sous-échantillons correspondants, tirons chaque grappe sélectionnée, commençons, puis, pour chaque grappe sélectionnée, tirons une unité au hasard dans le sous-échantillon correspondant. Il apparaît que l'échantillonnage inverse de premier degré de grappes ne soit pas nécessaire. Pour le montrer, notons que

$$p(s_0) = \left[\prod_{k=1}^K \left(\frac{N}{M_i^t} \right) \right] \left[\prod_{i=1}^k \frac{1}{M_{i'}^t} \right] = \left(\frac{1}{N} \right)^k$$

où $m_{i'}^t$ est la taille de sous-échantillon associée à la grappe sélectionnée lors du i^{e} tirage ($i = 1, \dots, k$). Nous voulons tirer un sous-échantillon s^* de taille k tel que

$$p(s^*) = (1/N)^k, \text{ où } N = \sum_{i=1}^K M_i^t. \text{ En outre, le nombre de termes dans } \sum_{s_0 \subset s^*} \text{ est égal à } \prod_{i=1}^k (M_i^t - 1) \text{ et}$$

$(\pi_i^*(s_0), \pi_i^h(s_0))$. Si les probabilités d'inclusion conditionnelles ne dépendent pas de s_0 , alors nous les écrivons sous la forme (π_i^*, π_i^h) . On voit aisément que

$$\pi_i^* = \sum_{s_0 \in I} p_0(s_0) \pi_i^*(s_0); \quad \pi_i^h = \sum_{s_0 \in I} p_0(s_0) \pi_i^h(s_0). \quad (2.3)$$

Si $\pi_i^*(s_0) = \pi_i^*$ et $\pi_i^h(s_0) = \pi_i^h$, alors il découle de (2.3) que

$$\pi_i^* = \pi_i^*, \pi_i^h = \pi_i^h. \quad (2.4)$$

À la section 4, nous utilisons (2.4) pour étudier les propriétés de l'échantillonnage inverse pour l'estimation d'un total de population. Notons que (π_i^*, π_i^h) peut correspondre à un autre plan de sondage plus simple s'il est impossible d'égaliser l'échantillonnage aléatoire simple (EAS), comme l'échantillonnage aléatoire simple stratifié.

2.1 Échantillonnage aléatoire simple stratifié

Supposons que l'échantillon original s_0 est un échantillon aléatoire simple stratifié, c'est-à-dire

$$p_0(s_0) = \prod_{L=1}^h \binom{N_h}{n_h}, \quad (2.5)$$

où $N_h(n_h)$ représente le nombre d'unités de population (d'échantillon) dans la strate h ($= 1, \dots, L$). Nous souhaitons tirer un sous-échantillon s^* de taille m , tel que $p(s^*) = 1 / \binom{m}{N_h}$, où $N = \sum_{h=1}^L N_h$. Manifestement, m ne peut pas être plus grand que $\min(n_h)$. Soit $\mathbf{m} = (m_1, \dots, m_L)$ le nombre (aléatoire) d'unités, qui, dans chaque strate, appartient à s^* , $0 \leq m_h \leq m$, $\sum_{h=1}^L m_h = m$. En notant que le nombre de termes dans $\sum_{s_0 \in s^*}$ est égal à

$$\prod_{h=1}^L \binom{N_h}{m_h - m_h^*}, \quad \text{il découle de (2.2) que}$$

$$p(s^* | s_0) = \frac{\binom{m}{N} \prod_{h=1}^L \binom{N_h}{m_h}}{\prod_{h=1}^L \binom{N_h}{m_h}}. \quad (2.6)$$

Le scénario de sous-échantillonnage se déduit facilement de (2.6) : i) générer \mathbf{m} à partir de la distribution hyper-géométrique $f(\mathbf{m}) = \prod_{h=1}^L \binom{N_h}{m_h} / \binom{N}{m}$; ii) tirer un échantillon aléatoire simple de taille m_h , sans remise, à partir des n_h unités échantillonnées dans la strate h , indépendamment dans les diverses strates h ($= 1, \dots, L$). HOS spécifient $p(s^* | s_0)$ pour commencer, puis vérifient qu'elle donne $p(s^*) = \binom{m}{N}^{-1}$. Notre approche fournit le scénario de sous-échantillonnage à partir de la spécification de $p_0(s_0)$ et $p(s^*)$.

2.2 Échantillonnage en grappes à un degré

HOS ont étudié le cas de l'échantillonnage en grappes à un degré en détail. Ils ont examiné trois plans d'échantillonnage pour s_0 , à savoir 1) grappes de taille égale, M , et échantillonnage simple des grappes; 2) grappes de taille inégale, M_i , et échantillonnage aléatoire simple des

indépendamment un grand nombre de fois et en calculant la moyenne des résultats.

Est-il possible de produire des sous-échantillons ayant les propriétés souhaitées? La réponse est souvent affirmative, quoique la taille du sous-échantillon résultant, m , doit parfois être petite (en fait, pas plus grande que $m = 2$ pour certains plans de sondage à plusieurs degrés stratifiés types). HOS donnent des algorithmes pour produire des échantillons inverses aléatoires simples pour un certain nombre de plans de sondage types. Nous résumons les scénarios d'échantillonnage inverse à la section 2 à titre de référence. Ces scénarios incluent des méthodes donnant une concordance exacte ou approximative avec l'échantillonnage aléatoire simple. Dans le présent article, nous examinons certaines propriétés des méthodes d'échantillonnage inverses répétées décrites à la section 2. En particulier, nous élaborons une théorie de l'échantillonnage inverse à la section 3 et illustrons certains de ses points forts et points faibles. À la section 4, nous étudions le cas spécial d'un total de population. À la section 5, nous proposons une approche fondée sur des équations d'estimation combinées (EEC) pour traiter les paramètres complexes comme les ratios et les paramètres de régression « en cas de recensement ». Enfin, nous présentons certaines conclusions à la section 6.

Les preuves de théorèmes sont présentées en annexe.

2. ALGORITHMES D'ÉCHANTILLONNAGE INVERSE

À la présente section, nous résumons les scénarios d'échantillonnage inverse proposés par Hinkins et coll. (1997) à titre de référence. Ces scénarios incluent des méthodes exactes ainsi qu'approximatives en ce qui concerne la concordance inconditionnelle avec l'échantillonnage aléatoire simple (EAS).

Supposons que nous ayons un échantillon s_0 d'observations tiré d'une population finie de taille N conformément à un plan de sondage complexe précis. Nous souhaitons tirer un sous-échantillon s^* de taille m à partir de s_0 tel que la probabilité inconditionnelle de s^* , $p(s^*)$, concorde avec celle de l'échantillonnage aléatoire simple $p(s^*) = 1 / \binom{m}{N}$, exactement ou approximativement. Nous avons

$$p(s^*) = \sum_{s_0 \in s^*} p_0(s_0) p(s^* | s_0), \quad (2.1)$$

où $p_0(s_0)$ est la probabilité de sélectionner s_0 et $p(s^* | s_0)$ est la probabilité conditionnelle de choisir s^* . Si $p(s^* | s_0)$ ne dépend pas de s_0 , alors il découle de (2.1) que

$$p(s^* | s_0) = \frac{\sum_{s_0 \in s^*} p_0(s_0)}{p(s^*)}. \quad (2.2)$$

Représentons les probabilités d'inclusion de premier et de deuxième ordres correspondant à s^* et à s_0 par (π_1^*, π_2^*) et (π_1^0, π_2^0) , respectivement, où $\pi_i^* = m/N$ et $\pi_i^0 = m(m-1)/(N(N-1))$, $i \neq 1$. Pareillement, reprenons les probabilités d'inclusion conditionnelles par

Défaire les structures des données d'enquête complexes : théorie élémentaire et applications de l'échantillonnage inverse

J.N.K. RAO, A.J. SCOTT ET E. BENHIN¹

RÉSUMÉ

L'application des méthodes statistiques classiques aux données provenant d'enquêtes complexes sans tenir compte des caractéristiques du plan de sondage peut donner lieu à des inférences incorrectes. Certaines méthodes ont été mises au point pour tenir compte du plan de sondage, mais elles nécessitent des renseignements supplémentaires, comme les poids de sondage, les effets de plan ou l'identification des grappes pour les microdonnées. L'échantillonnage inverse (Hinkins, Oh et Scheuren 1997) offre une autre approche qui consiste à défaire les structures des données d'enquête complexe de sorte qu'on puisse appliquer les méthodes classiques. Des sous-échantillons répétés sont tirés selon un plan d'échantillonnage aléatoire simple indépendant et analysés individuellement par les méthodes types, puis combinés pour augmenter l'efficacité. Cette méthode permet de préserver le caractère confidentiel des microdonnées, mais elle nécessite une grande capacité de calcul. Nous présentons une théorie de l'échantillonnage inverse et explorons ses limites. Nous proposons une approche fondée sur des équations d'estimation combinées pour traiter les paramètres complexes, tels que les ratios et les paramètres de régression linéaire ou logistique « en cas de recensement ». La méthode est appliquée à un ensemble de données corrélées en grappes présentées dans Bataese, Harter et Fuller (1988).

MOTS CLÉS : Équations d'estimation combinées; confidentialité; sous-échantillonnage répété.

1. INTRODUCTION

La distinction entre le point de concentration de la méthodologie d'enquête conventionnelle et celui du reste de la statistique appliquée est assez nette. Les spécialistes de l'échantillonnage se sont attachés à la recherche de moyens efficaces (mais compliqués) de tirer des échantillons pour estimer des quantités assez simples (moyennes, proportions, totaux, etc., de population). En revanche, la plupart des autres spécialistes de la statistique appliquée se sont concentrés sur l'élaboration de méthodes complexes permettant l'ajustement de modèles très compliqués, mais appuyant sur un schéma hypothétique d'échantillonnage assez simple (souvent, l'indépendance des observations). En réalité, les données provenant d'enquêtes complexes sont souvent utilisées pour ajuster des modèles compliqués. Par exemple, certaines personnes voudraient vouloir utiliser des données provenant de l'Enquête sur la population active pour caractériser l'association entre les niveaux de scolarité et de chômage. D'autres voudraient souhaiter se servir des données provenant des enquêtes sur la santé pour étudier l'association entre les conditions de logement ou la pauvreté et la morbidité, et ainsi de suite. Étendre le champ d'application des méthodes standard afin de pouvoir les appliquer aux données provenant d'enquêtes complexes basées sur un échantillonnage à plusieurs degrés et des probabilités de sélection variables est un exercice difficile et fastidieux; voir, par exemple Skinner, Holt et Smith (1989). Comment les praticiens surmontent-ils les problèmes que pose la complexité de la structure des données? Hinkins,

Oh et Scheuren (1997) (appelés ci-après HOS) ont affirmé que si votre seul est un martéau, chaque problème a l'air d'un clou! Un des grands projets statistiques (SAS, Splus, SPSS, etc.). La plupart d'entre eux se limitent à faire passer leurs données dans un programme type et ne tiennent pas compte des caractéristiques du plan de sondage. Et ce, bien qu'on ait consacré, au cours des deux dernières décennies, beaucoup d'énergie à l'élaboration de méthodes d'analyse des données d'enquête qui tiennent compte des caractéristiques du plan de sondage et que des programmes spécialisés, comme SUDAAN ou WesVar, existent maintenant pour appliquer certaines de ces méthodes. Au lieu de développer de nouveaux outils complexes (qui ne seraient peut-être utilisés que rarement en pratique de toute façon), on peut travailler à rebours : plutôt que d'adapter les méthodes aux données, on peut adapter les données aux méthodes. Une approche de ce genre a été élaborée par Rao et Scott (1992; 1999). Une autre a été proposée dans HOS. Leur idée fondamentale est d'éviter les difficultés que cause un échantillon complexe en choisissant un sous-échantillon (échantillon inverse) dont la structure est inconditionnellement celle d'un échantillon aléatoire simple (ou du moins une structure considérablement plus simple à manipuler que celle de l'échantillon original). Naturellement, cette approche entraîne une perte d'efficacité, particulièrement si le sous-échantillon est beaucoup plus petit que l'échantillon original, comme cela s'avère souvent nécessaire. Nous pouvons toutefois augmenter l'efficacité en répétant le processus

¹ J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Canada, K1A 5B6. Courriel : jrao@math.carleton.ca; A.J. Scott, Department of Statistics, University of Auckland, Auckland, Nouvelle-Zélande. Courriel : scott@stat.auckland.ac.nz; E. Benhin, Division des méthodes d'enquête auprès des ménages, Statistique Canada, Ottawa, Canada, K1A 0T6. Courriel : emmanuel.benhin@statcan.ca.

Reiter envisage le problème de l'inférence pour des ensembles de microdonnées partiellement synthétiques. Les organismes statistiques peuvent diffuser des ensembles de microdonnées entièrement synthétiques afin d'assurer le respect de la confidentialité des renseignements fournis par les répondants. Bien qu'on ait mis au point des méthodes d'inférence applicables à un ensemble de données entièrement synthétiques, la plupart des organismes ne diffusent que des ensembles de données partiellement synthétiques, c'est-à-dire des ensembles de données pour lesquels seules les valeurs des variables délicates sont imputées. Peu de travaux ont été publiés à ce sujet. Reiter montre que la méthode qu'il propose est valide dans un cadre d'analyse bayésien et dans un cadre d'analyse fondé sur le plan de sondage, et illustre cette méthode grâce à des études par simulation.

Brewer et Donadio décrivent l'obtention d'un estimateur de la variance de l'estimateur de Horvitz-Thompson qui ne nécessite pas le calcul des probabilités d'inclusion de deuxième ordre dans des conditions de grande entropie. Ces conditions surviennent lorsqu'il n'existe aucune régularité ni aucun ordonnancement décelable dans les unités échantillonnées. Dans des conditions de grande entropie, ils établissent une formule de la variance approximative et la vérifient par une approche assistée par modèle. Puis, ils élaborent un estimateur de cette variance approximative basé sur le plan de sondage pour l'estimateur de Horvitz-Thompson. Enfin, ils comparent empiriquement l'estimateur proposé à plusieurs autres estimateurs qu'ils appliquent à plusieurs populations.

Enfin, Chow et Thompson présentent une approche bayésienne des plans de sondage où les liens sociaux sont exploités pour échantillonner des populations humaines cachées ou d'accès difficile. Ils donnent une introduction facile à comprendre à l'approche bayésienne dans laquelle les liens sociaux entre les personnes sont utilisés pour créer les lois de distribution a priori. Il est facile d'ajuster ces lois a priori lorsque l'information est vague. Par conséquent, les lois a posteriori résultantes permettent de répondre à un grand nombre de questions.

M.P. Singh

Dans ce numéro

Le présent numéro de *Techniques d'enquête* comprend une section spéciale sur l'erreur de couverture dans les recensements comportant quatre articles, y compris deux sur l'enquête d'évaluation de la couverture utilisée aux États-Unis, un sur celle réalisée en Turquie et un sur celle réalisée en Italie. Cette section spéciale est précédée d'un article accompagné d'une discussion et suivie de quatre articles traitant de sujets divers.

Dans le premier article, Rao, Scott et Benhin étudient la méthode d'échantillonnage inverse répétée proposée par Hinkins, Oh et Scheuren. Selon cette approche, des sous-échantillons aléatoires sont sélectionnés à partir d'un échantillon complexe de façon à ce que chaque sous-échantillon soit inconditionnellement un échantillon aléatoire simple tiré de la population. Rao, Scott et Benhin présentent certains résultats théoriques concernant l'espérance et la variance de l'estimateur par échantillonnage inverse répété, puis ils examinent certaines conditions sous lesquelles cet estimateur converge vers l'estimateur original sur échantillon complet. Enfin, ils proposent une approche fondée sur les équations d'estimation permettant d'éviter certains biais éventuels de l'estimateur sous échantillonnage inverse répété pour les paramètres non linéaires. Cet article est suivi de deux discussions fascinantes, l'une rédigée par Etinge et l'autre par Hinkins, et d'une réplique des auteurs.

Dans le premier article de la section spéciale sur les erreurs de couverture dans les recensements, Hogan présente un aperçu concis de l'enquête utilisée pour produire des estimations du sous-dénombrement net au Recensement de 2000. Il présente l'étude d'évaluation de l'exactitude et de la couverture (ACE pour Accuracy and Coverage Evaluation) dans le contexte d'enquêtes postcensitaires générales et des estimateurs à système dual. Il présente aussi les hypothèses nécessaires, dans le cas de ces types d'enquête, pour produire des estimations sans biais, ainsi qu'une discussion détaillée des situations où ces hypothèses n'ont pas été vérifiées dans le cas de l'ACE de 2000. Les résultats sont très intéressants.

L'article suivant traite aussi de l'ACE de 2000. Cantwell et Ikeda examinent les hypothèses cruciales qui sont émise lorsque certaines données manquent. L'un des points qu'ils soulignent est que, si l'on estime une caractéristique rare, comme le fait de ne pas être dénombré lors du recensement, le choix de la méthode utilisée pour faire la correction pour les données manquantes est très important. Les auteurs décrivent les modifications apportées aux méthodes utilisées lors des enquêtes postcensitaires qui ont précédé l'ACE de 2000.

Ayhan et Eknî présentent les procédures d'évaluation de la couverture utilisées dans un contexte de recensement différent. Bien que la Turquie utilise le plan d'enquête postcensitaire de base, il existe des différences intéressantes entre les procédures utilisées dans ce pays et celles utilisées aux États-Unis. Puisque la Turquie suit une approche *de facto* pour déterminer le lieu de résidence au moment du recensement plutôt que l'approche *de jure* utilisé aux États-Unis, les enquêtes postcensitaires présentent certaines différences opérationnelles qui sont décrites clairement par les auteurs.

Le dernier article de la section spéciale sur les erreurs de couverture dans les recensements, rédigé par Cocchi, Fabrizio et Trivisano, décrit le Recensement de la population de l'Italie de 1991 et l'Enquête postcensitaire (EPC) utilisée pour évaluer le sous-dénombrement. Puisque le recensement est administré par les municipalités, les données sur la qualité statistique des municipalités sont utilisées comme données auxiliaires pour la modélisation et l'estimation fondées sur les données de l'EPC. Les auteurs utilisent des arbres de régression de Poisson et des modèles hiérarchiques de Poisson pour analyser les données. Ils résument leurs résultats et en discutent, et font certaines recommandations.

Skinner et Carter étendent l'estimation de la mesures du risque de divulgation applicable aux microdonnées d'enquête élaborée par Skinner et Elliot de l'échantillonnage avec probabilités égales à l'échantillonnage avec probabilités inégales sous l'hypothèse d'un échantillonnage de Poisson. Ils considèrent aussi les effets des écarts éventuels par rapport à cet échantillonnage.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 29, numéro 2, décembre 2003

TABLe DES MATIÈRES

Dans ce numéro	117
Article de discussion	
J.N.K. RAO, A.J. SCOTT et E. BENHIN	119
Défaite les structures des données d'enquête complexes : théorie élémentaire et applications	119
de l'échantillonnage inverse	133
Commentaires : JOHN L. ELTINGE	136
SUSAN HINKINS	140
Réponse des auteurs	140
Section spéciale sur les erreurs de couverture dans les recensements	
HOWARD HOGAN	145
L'évaluation de l'exactitude et de la couverture : théorie et conception	145
PATRICK J. CANTWELL et MICHAEL IKEDA	157
Traitement des données manquantes dans l'enquête d'évaluation de l'exactitude	157
et de la couverture de 2000	157
H. ÖZTAŞ AYHAN et SÜHENDAN EKNİ	175
Erreur de couverture des recensements de population : Le cas de la Turquie	175
D. COCCHI, E. FABRIZI et C. TRIVISANO	187
Un modèle hiérarchique pour l'analyse du sous-dénombrement local du recensement en Italie	187
Articles Réguliers	
C.J. SKINNER et R.G. CARTER	197
Estimation d'une mesure du risque de divulgation pour les microdonnées d'enquête sous échantillonnage	197
avec probabilités inégales	197
J.P. REITER	203
Inference pour les ensembles de microdonnées à grande diffusion partiellement synthétiques	203
K.R.W. BREWER et MARTIN E. DONADIO	213
La variance sous grande entropie de l'estimateur de Horvitz-Thompson	213
MOSUK CHOW et STEVEN K. THOMPSON	221
Estimation avec plans d'échantillonnage par dépistage de liens - Une approche bayésienne	221
Remerciements	231

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président	G.J. Brackstone
Membres	D.A. Binder G.J.C. Hole C. Patrick R. Platek (Ancien président)
COMITÉ DE DIRECTION	

COMITÉ DE RÉDACTION

Rédacteur en chef	M.P. Singh, <i>Statistique Canada</i>
Rédacteurs associés	D.R. Bellhouse, <i>University of Western Ontario</i> D.A. Binder, <i>Statistique Canada</i> J.M. Brick, <i>Westat, Inc.</i> C. Clark, <i>U.S. Bureau of the Census</i> J. Eltinge, <i>U.S. Bureau of Labor Statistics</i> W.A. Fuller, <i>Iowa State University</i> J. Gambino, <i>Statistique Canada</i> M.A. Hidiroglou, <i>Statistique Canada</i> G. Kalton, <i>Westat, Inc.</i> P. Kott, <i>National Agricultural Statistics Service</i> P. Lahiri, <i>JPSM, University of Maryland</i> S. Linacre, <i>Official National Statistics</i> G. Nathan, <i>Hebrew University, Israel</i>

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception de coulant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte préparé selon les directives présentes dans la revue et rédigé en anglais ou en français au rédacteur en chef, M. P. Singh, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Pré Tunney, Ottawa (Ontario), Canada K1A 0T6. Courriel : singhmp@statcan.ca. On peut aussi envoyer quatre exemplaires imprimés. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de *Techniques d'enquête* (n° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada. États-Unis 12 \$ CA (6 \$ x 2 exemplaires); autres pays, 30 \$ CA (15 \$ x 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commander par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiens et statisticiens du Québec.

Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Janvier 2004

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistré ou non, par quelque autre, ou de l'emmagasiner dans un système de recouvrement, magnétique, reproduction électronique, mécanique, photographique, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

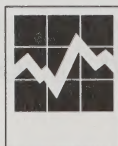
© Ministre de l'Industrie, 2004

Publication autorisée par le ministre
responsable de Statistique Canada

DÉCEMBRE 2003 • VOLUME 29 • NUMÉRO 2

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE





NUMÉRO 2

•

VOLUME 29

•

DÉCEMBRE 2003

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

N° 12-001-XPB au catalogue

TECHNIQUES D'ENQUÊTE

